Estimating the Kullback-Liebler risk based on multifold cross-validation

Paolo Vidoni
Department of Economics and Statistics, University of Udine
via Tomadini 30/a, I-33100 Udine, Italy
paolo.vidoni@uniud.it

Abstract

This paper concerns a class of model selection criteria based on cross-validation techniques and estimative predictive densities. Both the simple or leave-one-out and the multifold or leave-m-out cross-validation procedures are considered. These cross-validation criteria define suitable estimators for the expected Kullback-Liebler risk, which measures the expected discrepancy between the fitted candidate model and the true one. In particular, we shall investigate the potential bias of these estimators, under alternative asymptotic regimes for m. The results are obtained within the general context of independent, but not necessarily identically distributed, observations and by assuming that the candidate model may not contain the true distribution. An application to the class of normal regression models is also presented and simulation results are obtained in order to gain some further understanding on the behaviour of the estimators.

Keywords: AIC, cross-validation, Kullback-Liebler information, likelihood, misspecification, predictive density.

Acknowledgements: The author thanks an anonymous referee for suggestions and comments which served to improve the paper.

Running head: KL risk estimation based on cross-validation.

1 Introduction

Let us suppose that the observations $y = (y_1, \ldots, y_n)$, $n \geq 1$, are available for selecting a model from a given collection of plausible statistical models. A class of useful model selection criteria is based on cross-validation techniques and suitable predictive densities. The idea behind cross-validation is simple. It involves a split of the data y into two parts. The first one contains $n - m$ data points, with $m \in \{1, \ldots, n-1\}$, and it is called training set, since it is used for fitting the model which is considered. The second one, containing the remaining m data points, is used for assessing the predictive ability of the model under consideration and it is called validation set. All the splits of the data have to be potentially considered. The simplest version of cross-validation consists of leaving out one observation at a time, that is $m = 1$, and it is called simple or leave-one-out cross-validation. Whenever $m > 1$ observations are left out at a time, we have a multifold or a leave-m-out cross-validation procedure.

Discussion on the properties and the applicability of simple cross-validation procedures, in various situations usually related to model selection, may be found in a number of papers; see, for example, Allen (1974), Stone (1974, 1977a, 1977b), Geisser (1975) and Efron (1983, 1986). The leave-m-out cross-validation methods, with $m > 1$, first appear in Geisser (1975) and some subsequent developments may be found, for example, in Herzberg and Tsukanov (1986). Papers by Li (1987), Zhang (1993) and Shao (1993, 1997) investigate, from a theoretical point of view, the asymptotic properties of various multifold cross-validation criteria, principally applied to the case of alternative linear regression models. With regard to non-asymptotic properties, few results are available (see Arlot, 2008) and they are essentially based on the comparison with the so called *oracle* model, namely the model one would choose if the distribution of the observations were known.

The popularity and the attractiveness of cross-validation methods come from the fact that they are based on a simple and intuitive idea and that they can be applied to a variety of different model selection problems. In particular, cross-validation is used in the regression framework, usually with regard to the variable selection problem, in order to estimate a suitable notion of prediction error with the aim of choosing the appropriate predictors (see, for example, Burman, 1989, and Picard and Cook, 1984). Further interesting applications may be found in the density estimation context with the objective of assessing the accuracy of alternative density estimators (see, for example, the recent contribution by Celisse, 2014).

An extensive review on cross-validation procedures for model selection is given by Arlot and Celisse (2010), where a general classification is presented with the aim of emphasizing the performances of the

cross-validation methods in various model selection frameworks. However, even if many applied and theoretical contributions on cross-validation can be found in the literature, there are many important issues not completely understood, regarding in particular the multifold procedures, which need to be considered in more detail.

In this paper, following Geisser and Eddy (1979), we shall consider general cross-validation methods for model selection based on frequentist predictive densities, with particular reference to the estimative predictive density. As in Konishi and Kitagawa (1996), Burnham and Anderson (2002) and Fujikoshi, Noguchi, Ohtaki and Yanagihara (2003), we use cross-validation criteria in order to define suitable estimators for the expected Kullback-Liebler information, and in particular for the Kullback-Liebler risk, which measures the discrepancy between the fitted candidate model and the true one. Burman (1989) addresses the problem of risk estimation in the regression framework and finds asymptotic expansions for the mean and the variance of risk estimators based on alternative multifold cross-validation techniques. An estimator, which is a good solution according to the bias-variance trade-off criterion, is proposed. Recent results concerning the potential bias of these estimators may be found in Fushiki (2011), Yanagihara, Tonda and Matsumoto (2006) and Yanagihara and Fujisawa (2012).

The contribution of the present paper concerns the computation of the first-order bias term of both the simple and the multifold cross-validation selection statistics, by assuming alternative asymptotic regimes for m. These results are obtained within the general context of independent, but not necessarily identically distributed, observations and by assuming that the alternative models may not necessarily contain the true distribution. Thus, the framework is usually different from those ones considered in the above mentioned papers. Moreover, parameter estimation is done by maximum likelihood inferential procedures and connections with the Akaike's information criterion are emphasized.

Besides the evaluation of the potential bias of the estimators, the asymptotic expansions derived in this paper may be also considered for defining asymptotic equivalent expressions for the cross-validation selection statistics. This can be extremely useful from the applied point of view, since the application of multifold cross-validation criteria is usually computationally demanding. Finally, an application of these results to model selection involving normal regression models is presented, giving results similar to those obtained by Zhang (1993) when the model contains the true distribution. A simple simulation study is also considered in order to gain some further understanding on the behaviour of these cross-validation estimators.

2 Assumptions and preliminary results

2.1 Preliminaries

Let us assume that the data $y = (y_1, \ldots, y_n)$, $n \geq 1$, are a realization of the random vector $Y = (Y_1, \ldots, Y_n)$, with an unknown joint probability density function $g(y)$, with respect to a suitable dominating measure. The marginal random variables $Y_1, \ldots, Y_n$ are supposed to be independent, so that we include both the simple case with independent, identically distributed, observations and the more general situation where the random variables are independent, but not identically distributed, such as within linear and generalized linear models. We consider as a plausible candidate statistical model for Y a parametric family of probability density functions, with respect to a common dominating measure, defined as $f(y;\omega)$, $\omega \in \Omega \subseteq \mathbf{R}^d$, $d \geq 1$, where ω is an unknown parameter. Since these family may not necessarily contain the true density $g(y)$, the model could be misspecified for Y. We assume that $f(y;\omega)$ is a smooth function, so that, for every fixed $\omega \in \Omega$, $f(y;\omega)$ is a measurable function in y and, for every fixed y, $f(y;\omega)$ is a function at least continuously differentiable of order two on Ω.

In a prediction-based model selection framework, the aim is to choose the model which offers the most satisfactory predictive explanation to the observed sample y. More precisely, we select the model which presents the best predictive ability, with respect to the available data y, considered as potential future observations. Under this respect, the expected Kullback-Liebler divergence between the true and the candidate models defines a suitable measure of the associated mean prediction error. Indeed, since this prediction error is in fact unknown, the best model is determined as that one minimizing a suitable estimator for the Kullback-Liebler risk. An interesting class of model selection criteria, which constitutes the focus of the present paper, is based on cross-validation estimators. Alternative loss functions could be considered instead of the Kullback-Liebler one, as emphasized in the review paper by Arlot and Celisse (2010). A popular loss function, frequently used in the regression context, is the quadratic loss function which, for linear Gaussian models, defines a model selection criterion equivalent to that one based on the Kullback-Liebler risk.

As emphasized in Section 1, cross-validation criteria require the split of the random vector Y into two disjoint parts: the first part plays the role of an observable random vector and the second one is viewed as a future random vector. The idea behind cross-validation is to validate the predictive ability of the candidate model, with respect to the observed data y, by means of the following general criterion.

Definition 2.1 *Let $Y = (Y_1, \ldots, Y_n)$ be a random vector as defined before and let q denote a subset of $\{1, \ldots, n\}$ of fixed size $m \in \{1, \ldots, n-1\}$. Let us define $Y_q = (Y_i,\, i \in q)$ and $Y_{(q)} = (Y_i,\, i \notin q)$, with y_q and $y_{(q)}$ the associated observed values. The cross-validation selection procedure points to the model maximizing the selection statistic*

$$\Psi_{CV(m)}(Y) = \frac{1}{\binom{n-1}{m-1}} \sum_{q} \log f(Y_q; \widehat{\omega}_{(q)}), \tag{2.1}$$

where $\widehat{\omega}_{(q)} = \widehat{\omega}(Y_{(q)})$ is the maximum likelihood estimator of ω based on $Y_{(q)}$. The summation is over all possible subsets q of size m.

Note that $\Psi_{CV(m)}(Y)$ involves the plug-in estimator of the density function $f(y_q; \omega)$, which corresponds to the estimative predictive density for Y_q obtained by substituting ω with $\widehat{\omega}_{(q)}$. Hereafter we consider maximum likelihood estimators for the parameter ω, even if any alternative asymptotic equivalent estimator for ω may be taken into account. Indeed, we shall adopt the notation $f(u; \omega)$ for the density of a random vector U, as indicated by the argument of function $f(\cdot; \omega)$.

The selection statistic (2.1), with $m > 1$, defines a multifold or a leave-m-out cross-validation procedure. Whenever $m = 1$, we have, as a particular case, a simple or a leave-one-out cross-validation procedure, with selection statistic given by $\Psi_{CV(1)}(Y)$. Geisser and Eddy (1979) call this simple selection procedure predictive sample reuse quasi-likelihood method. Although the distinction between $m = 1$ and $m > 1$ has an intuitive motivation, we will show in Section 3 that the asymptotic properties of $\Psi_{CV(m)}(Y)$, with a fixed m, not increasing with n, are equivalent to those of $\Psi_{CV(1)}(Y)$. Thus, a more realistic distinction would be between cross-validation procedures with m fixed and cross-validation procedures with m increasing with n.

Furthermore, we assume that for each $n \geq 2$, $m \in \{1, \ldots, n-1\}$ is such that $n - m = O(n^{\gamma})$ and $m = O(n^{\delta})$, as $n \to +\infty$, with $\gamma \in (0, 1]$ and $\delta \in [0, 1]$. Consequently, we state that both the case with a fixed m not increasing with n, which includes the simple cross-validation procedure, and the case with m increasing with n may be considered in this framework. If m is fixed, $\gamma = 1$ and $\delta = 0$. Note that we exclude the value $\gamma = 0$, which corresponds to the situation with $n - m$ not increasing with n. On the other hand, it is possible to have $\gamma = 1$ and $\delta = 1$. This happens, for example, when $m/n = \lambda + o(1)$, as $n \to +\infty$, with $\lambda \in (0, 1)$; that is, when the observations left out constitute nearly a fixed proportion λ of the whole sample. Moreover, $\widehat{\omega}$ corresponds to $\widehat{\omega}_{(q)}$, with $m = 0$.

In the sequel, we shall emphasize the properties of these general model selection criteria, viewed as estimators of the expected Kullback-Liebler divergence, with particular attention to the potential

bias of the inferential procedure.

2.2 Basic assumptions

Let us consider the maximum likelihood estimators for ω under the candidate model. In particular, we consider the estimators $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ based, respectively, on the full sample Y and on the sub-sample $Y_{(q)}$ defined as

$$\begin{aligned} \widehat{\omega} &= \arg\max_{\omega} \ell(\omega; Y) = \arg\max_{\omega} \sum_{i=1}^{n} \ell(\omega; Y_i), \\ \widehat{\omega}_{(q)} &= \arg\max_{\omega} \ell(\omega; Y_{(q)}) = \arg\max_{\omega} \sum_{i \notin q} \ell(\omega; Y_i), \end{aligned}$$

with $\ell(\omega; Y_i) = \log f(Y_i; \omega)$, $\ell(\omega; Y) = \log f(Y; \omega)$ and $\ell(\omega; Y_{(q)}) = \log f(Y_{(q)}; \omega)$ the associated loglikelihood functions.

In order to achieve the asymptotic results presented in the following sections, we implicitly require the regularity assumptions for the existence and the validity of the standard likelihood asymptotic results, under potentially misspecified statistical models (see, for example, White, 1994, Chapters 3 and 6). In particular, let us consider the sequence of values $\{\omega_n^*\}_{n \geq 1}$, called pseudo-true parameter values, minimizing with respect to ω the Kullback-Leibler divergence between $g(y)$ and $f(y; \omega)$

$$I_n(g, f; \omega) = E_Y \left\{ \log \frac{g(Y)}{f(Y; \omega)} \right\} = E_Y\{\log g(Y)\} - E_Y\{\log f(Y; \omega)\},$$

where the expectations are with respect to the true distribution of Y. According to White (1994, Definition 3.3), this sequence is required to be unique and such that $\omega_n^* \in int(\Omega)$ uniformly in n. These values may also be viewed as the maximizers of $E_Y\{\log f(Y; \omega)\}$, $n \geq 1$. Note that, if $Y_1, \ldots, Y_n$ are identically distributed, we have that $\omega_n^* \equiv \omega^*$, where $\omega^* \in int(\Omega)$ is the minimizer of the Kullback-Leibler divergence between the marginal densities $g(y_i)$ and $f(y_i; \omega)$, $i = 1, \ldots, n$.

Hereafter, we consider the special case where the maximum likelihood estimators $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ are defined as the solutions to the equations

$$\sum_{i=1}^{n} \partial_r \ell(\widehat{\omega}; Y_i) = 0, \quad r = 1, \ldots, d, \tag{2.2}$$

$$\sum_{i \notin q} \partial_r \ell(\widehat{\omega}_{(q)}; Y_i) = 0, \quad r = 1, \ldots, d, \tag{2.3}$$

where $\partial_r \ell(\widehat{\omega}; Y_i)$ and $\partial_r \ell(\widehat{\omega}_{(q)}; Y_i)$, $i = 1, \ldots, n$, are $\partial_r \ell(\omega; Y_i) = \partial \ell(\omega; Y_i)/\partial \omega_r$ evaluated at $\omega = \widehat{\omega}$ and $\omega = \widehat{\omega}_{(q)}$, respectively. Indeed, ω_r, $r = 1, \ldots, d$, is the r-th component of vector ω. Moreover, the

pseudo-true parameter values $\{\omega_n^*\}_{n\geq 1}$ are such that, for $n \geq 1$,

$$E_Y\{\partial_r \ell(\omega_n^*; Y)\} = 0, \quad r = 1, \ldots, d, \tag{2.4}$$

where $\partial_r \ell(\omega_n^*; Y)$ is $\partial_r \ell(\omega; Y) = \partial \ell(\omega; Y)/\partial \omega_r$ evaluated at $\omega = \omega_n^*$.

We shall restrict our attention to statistical models where the maximum likelihood estimators $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ are weakly consistent for ω_n^* and ω_{n-m}^*, respectively; that is, $\widehat{\omega} - \omega_n^* \to 0$ in probability, as $n \to +\infty$, such that $\widehat{\omega} - \omega_n^* = O_p(n^{-1/2})$, and analogously for $\widehat{\omega}_{(q)}$, with $\widehat{\omega}_{(q)} - \omega_{n-m}^* = O_p(n^{-\gamma/2})$. Moreover, under proper conditions, as specified in White (1994, Chapter 6), $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ are asymptotically normally distributed. The asymptotic covariance matrix of $\widehat{\omega}$ is such that

$$E_Y\{(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)\} = \nu_{t,u} i^{rt} i^{us} + o(n^{-1}), \quad r, s = 1, \ldots, d, \tag{2.5}$$

where $\widehat{\omega}_r$ and ω_r^*, $r = 1, \ldots, d$, are the r-th components of vectors $\widehat{\omega}$ and ω_n^*, respectively, with n omitted to simplify the notation. Here, $\nu_{r,s} = E_Y\{\partial_r \ell(\omega_n^*; Y) \partial_s \ell(\omega_n^*; Y)\}$ and $\nu_{rs} = E_Y\{\partial_{rs} \ell(\omega_n^*; Y)\}$, where $\partial_{rs} \ell(\omega_n^*; Y)$ is $\partial_{rs} \ell(\omega; Y) = \partial \ell(\omega; Y)/\partial \omega_r \partial \omega_s$, evaluated at $\omega = \omega_n^*$. Indeed, i^{rs} is the (r, s)-element of the inverse of the expected information matrix $[i_{rs}] = [-\nu_{rs}]$. An analogous result holds for $\widehat{\omega}_{(q)}$, with an error term of order $o(n^{-\gamma})$. In order to derive formula (2.5), we consider the fact that

$$\partial_{rs} \ell(\omega_n^*; Y) - \nu_{rs} = O_p(n^{1/2}), \quad r, s = 1, \ldots, d. \tag{2.6}$$

Hereafter we adopt the Einstein summation convention so that, whenever an index appears more than once in a single term, summation over that index is understood. The convention is suppressed for indices related to the observations.

In the particular case when the assumed model is correctly specified, that is when $g(y) = f(y; \omega_0)$, for some $\omega_0 \in int(\Omega)$, we have that $I_n(g, f; \omega)$, $n \geq 1$, attains its minimum at $\omega = \omega_0$ and then $\omega_n^* \equiv \omega_0$, with ω_0 the true parameter value. Moreover, the well-known information identity $\nu_{rs} = -\nu_{r,s}$ holds and we obtain the usual formula

$$E_Y\{(\widehat{\omega}_r - \omega_{0r})(\widehat{\omega}_s - \omega_{0s})\} = i^{rs} + o(n^{-1}), \quad r, s = 1, \ldots, d,$$

with ω_{0r} the r-th component of vector ω_0.

2.3 Expansions for maximum likelihood estimators

Since the cross-validation procedures, considered in this paper, require the computation of the maximum likelihood estimators based on $Y_{(q)}$, with different choices for q, it can be useful, both from the

theoretical and the computational point of view, to investigate the asymptotic relationship between $\widehat{\omega}_{(q)}$ and $\widehat{\omega}$.

Proposition 2.1 *Under the assumptions stated in Section 2.2, if $\delta \neq 1$, the maximum likelihood estimators $\widehat{\omega}_{(q)}$ and $\widehat{\omega}$ are such that*

$$\widehat{\omega}_{(q)r} = \widehat{\omega}_r + \partial_s \ell(\widehat{\omega}; Y_q)\partial^{rs}\ell(\widehat{\omega}; Y) + o_p(n^{\delta-1}), \quad r = 1, \ldots, d, \tag{2.7}$$

where $\partial_s \ell(\widehat{\omega}; Y_q) = \sum_{i \in q} \partial_s \ell(\widehat{\omega}; Y_i)$, $\partial^{rs}\ell(\widehat{\omega}; Y)$ is the (r, s)-element of the inverse of matrix $[\partial_{rs}\ell(\widehat{\omega}; Y)]$ and $\widehat{\omega}_{(q)r}$ is the r-th components of vector $\widehat{\omega}_{(q)}$.

The proof of Proposition 2.1 is deferred to the Appendix. Notice that if m is fixed, then $\gamma = 1$, $\delta = 0$ and it is easy to see that $\widehat{\omega}_{(q)} - \widehat{\omega} = O_p(n^{-1})$. In particular, when $m = 1$ and $q = \{i\}$, $i \in \{1, \ldots, n\}$, (2.7) particularizes to

$$\widehat{\omega}_{(q)r} = \widehat{\omega}_r + \partial_s \ell(\widehat{\omega}; Y_i)\partial^{rs}\ell(\widehat{\omega}; Y) + o_p(n^{-1}), \quad r = 1, \ldots, d.$$

Furthermore, it can be useful to emphasize that if $\delta = 1$, then $m = O(n)$ and a result similar to (2.7) can not be achieved. The point is that the main term and the remainder in the right hand side of (A.2), specified in the proof of Proposition 2.1, have the same asymptotic order and they both have to be considered in the expansion for $\partial_r \ell(\widehat{\omega}_{(q)}; Y_q)$. Thus, the inversion procedure can not be applied and a simple explicit relation linking $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ is not available.

Finally, equation (2.7) may be useful in the calculation of the cross-validation criterion, in order to reduce the computational burden, since it is possible to define $\widehat{\omega}_{(q)}$ as a suitable function of $\widehat{\omega}$, eventually with an error term of a low asymptotic order. Thus, the evaluation of the selection statistic (2.1) will require only one fit of the model, instead of the $\binom{n}{m}$ fits needed in principle.

3 Asymptotic bias of cross-validation criteria based on the estimative predictive density

3.1 Predictive information criterion for model selection

In this section, adopting the framework previously introduced, we shall study the cross-validation selection statistic (2.1), viewed as a suitable estimator for a target value measuring the predictive ability of a given parametric statistical model for the random vector Y. Following Konishi and

Kitagawa (1996) and Burnham and Anderson (2002), the goodness of a given model may be judge by considering the following indicator based on the Kullback-Leibler divergence.

As mentioned in Section 2.1, alternative divergences or loss functions may be considered for evaluating the predictive ability of a given model or, more generally, of a given statistical algorithm (see Arlot and Celisse, 2010). The choice of the loss function depends on the objective of the selection procedure and on the particular type of models or statistical algorithms taken into account. In this paper we focus on the Kullback-Leibler divergence and the associated logarithmic score function, which are usually considered whenever the aim is to select a density function viewed as a suitable estimator for the true density. The AIC-type model selection criteria, and more generally those ones based on modifications of the maximized loglikelihood, are usually defined in this framework.

Definition 3.1 *Let us consider an observable random vector $Y = (Y_1, \ldots, Y_n)$ and a future random vector $Z = (Z_1, \ldots, Z_n)$, with the same distribution as Y; Y and Z are supposed to be independent. Under the assumptions stated in Section 2, the expected Kullback-Leibler divergence (Kullback-Leibler risk) between the true density for Z, given by $g(z)$, and the estimative predictive density $\widehat{f}(z) = f(z;\widehat{\omega})$ under the candidate model is*

$$E_Y\{I_n(g, \widehat{f}; \omega)\} = E_Y[E_Z\{\log g(Z)\}] - E_Y[E_Z\{\log f(Z;\widehat{\omega})\}], \tag{3.1}$$

where the expectations are with respect to the common true distribution of Y and Z, $\widehat{\omega} = \widehat{\omega}(Y)$ and $I_n(g, \widehat{f}; \omega)$ is the Kullback-Liebler divergence between $g(z)$ and $f(z;\widehat{\omega})$.

The associated selection criterion, which points to the model minimizing the expected Kullback-Leibler risk (3.1), is equivalent to that one selecting the model maximizing the expected predictive loglikelihood

$$\eta(g, \widehat{f}) = E_Y[E_Z\{\log f(Z;\widehat{\omega})\}].$$

Note that, if the components of vectors Y and Z are independent and identically distributed, $\eta(g, \widehat{f}) = nE_Y[E_{Z_i}\{\log f(Z_i;\widehat{\omega})\}]$, namely n times the expected predictive loglikelihood for the marginal random variable Z_i, $i = 1, \ldots, n$. Konishi and Kitagawa (1996) consider this simple case, using functional-type estimators for ω.

Since the expected predictive loglikelihood depends on the true unknown distribution, we aim to define, as a model selection statistic, a suitable estimator $\Psi(Y)$ for $\eta(g, \widehat{f})$. In particular, we shall study the potential bias of these estimators for $\eta(g, \widehat{f})$, focusing on the situation where, exactly or

approximately, $E_Y\{\Psi(Y)\} = \eta(g, \widehat{f})$. Under this respect, the following proposition may be useful, since it provides asymptotic approximations for the target value $\eta(g, \widehat{f})$ and for the expectation of $\ell(\widehat{\omega}; Y) = \log f(Y; \widehat{\omega})$, which is the maximized loglikelihood associated to the candidate model. In the following, we usually consider ω^* instead of ω_n^*, omitting n in order to simplify the notation.

Proposition 3.1 *Under the assumptions stated before, we have that*

$$\eta(g, \widehat{f}) = E_Y\{\log f(Y; \omega^*)\} - \frac{1}{2}\,\nu_{t,r} i^{rt} + O(n^{-1}), \tag{3.2}$$

$$E_Y\{\ell(\widehat{\omega}; Y)\} = E_Y\{\log f(Y; \omega^*)\} + \frac{1}{2}\,\nu_{t,r} i^{rt} + O(n^{-1}). \tag{3.3}$$

The proof of Proposition 3.1 is deferred to the Appendix. A comparison between (3.2) and (3.3) shows that, to the relevant order of approximation, $\ell(\widehat{\omega}; Y)$ is usually an upwardly biased estimator for $\eta(g, \widehat{f})$. A motivation is related to the fact that, in this case, the same data are used to fit the model and to asses its predictive ability. A first order bias-corrected modification of the maximized loglikelihood is $\ell(\widehat{\omega}; Y) - \nu_{t,r} i^{rt}$. In the particular case with independent, identically distributed, observations, we obtain that $E_Y\{\log f(Y; \omega^*)\} = n\, E_{Y_i}\{\log f(Y_i; \omega^*)\}$ and $\nu_{t,r} i^{rt} = \bar{\nu}_{t,r} \bar{\imath}^{rt}$, where the expected likelihood quantities $\bar{\nu}_{t,u}$, $\bar{\imath}^{rt}$ are related to a single component Y_i. If the model contains the true distribution, $\nu_{t,r} i^{rt} = d$ and the modification of the maximized loglikelihood corresponds to the AIC, namely the Akaike Information Criterion (Akaike, 1973).

3.2 Simple and multifold cross-validation procedures

In this section we shall consider simple and multifold cross-validation procedures, based on the estimative predictive density, with selection statistic $\Psi_{CV(m)}(Y)$ given by (2.1), with $m \geq 1$. An interesting interpretation for this selection statistic may be given by the following equivalent expression

$$\Psi_{CV(m)}(Y) = \sum_{i=1}^{n} \frac{1}{\binom{n-1}{m-1}} \sum_{b(i)} \log f(Y_i; \widehat{\omega}_{b(i)}), \tag{3.4}$$

where $b(i)$ denotes a subset of $\{1, \ldots, i-1, i+1, \ldots, n\}$ of dimension $n - m$ and $\widehat{\omega}_{b(i)} = \widehat{\omega}(Y_{b(i)})$ is the maximum likelihood estimator for ω based on $Y_{b(i)} = (Y_j, j \in b(i))$. The summation is over all the $\binom{n-1}{n-m} = \binom{n-1}{m-1}$ subsets $b(i)$ of size $n - m$. Thus, $\Psi_{CV(m)}(Y)$ may be viewed as an estimator for $\eta(g, \widehat{f}) = \sum_{i=1}^{n} E_Y[E_{Z_i}\{\log f(Z_i; \widehat{\omega})\}]$, given by the sum of suitable estimators for each term $E_Y[E_{Z_i}\{\log f(Z_i; \widehat{\omega})\}]$, $i = 1, \ldots, n$, based on the $\binom{n-1}{n-m}$ samples $b(i)$ of size $n-m$. Note that, if $m = 1$,

each estimator involves a single sample $b(i)$ of dimension $n-1$ and, as expected, (3.4) corresponds to $\Psi_{CV(1)}(Y)$.

It is known that, since cross-validation procedures separate the data used to fit the model and the data used to define the prediction rule, the selection statistic (2.1) does not lead to substantial overfitting. With regard to the choice between simple and multifold criteria, an intuitive motivation for using $\Psi_{CV(m)}(Y)$, with $m > 1$, is that, leaving out groups of observations, rather than single observations, may avoid the potential problems of high variability related to $\Psi_{CV(1)}(Y)$. This statement may be motivated by comparing the case $m = 1$ and the case $m > 1$, with regard to the alternative formulation for $\Psi_{CV(m)}(Y)$ given by relation (3.4). However, as noted by Davison and Hinkley (1997) for linear models, multifold cross-validation procedures present a reduced variance but, usually, at the cost of an increasing bias.

In the sequel, we shall investigate the asymptotic bias of cross-validation criteria based on the estimative predictive density, viewed as estimators for the expected predictive loglikelihood $\eta(g, \widehat{f})$. It is well-known that a number of papers concerns the study of the potential bias of model selection statistics and proposes various bias-corrected solutions. With regard to cross-validation criteria, Burman (1989) considers the v-fold cross-validation criterion and the repeated learning-testing method as less computationally demanding alternatives to the simple cross-validation. He specifies a target quantity which generalizes $\eta(g, \widehat{f})$ and derives useful first-order expansions for both the mean and the variance of the estimators taken into account. Finally, he proposes a solution which accounts for the bias-variance trade-off and it is less computational expensive than the simple cross-validation. Furthermore, recent contributions by Fushiki (2011), Yanagihara, Tonda and Matsumoto (2006) and Yanagihara and Fujisawa (2012) adopt a framework quite similar to that one considered in the present paper and aim at defining first-order bias corrected cross-validation criteria.

The following two theorems, which represent the main original contribution of this paper, derive suitable asymptotic approximations for the expectation of $\Psi_{CV(m)}(Y)$, with respect to the true distribution of Y. The first one, with the associated corollary, concerns the case where $\delta \neq 1$ and it includes the simple cross-validation and the situation with $m > 1$ fixed. The second one regards the case where $\delta \in (0, 1]$, $\gamma \in (1/2, 1]$ and it includes the case where $\delta = \gamma = 1$. These results, although focussed only on the bias of the estimators, are obtained by assuming alternative asymptotic regimes for m and emphasize some interesting theoretical features of cross-validation criteria, not yet considered in the literature, useful as well for applications.

Theorem 3.1 *Under the assumptions stated before, if $\delta \neq 0, 1$, the selection statistic (2.1) is such that*

$$\begin{aligned} E_Y\{\Psi_{CV(m)}(Y)\} &= E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\,\nu_{t,r}i^{rt} - \frac{1}{\binom{n-1}{m-1}}\sum_q \nu_{q;r,s}i^{rs} \\ &+ \frac{1}{2}\,\frac{1}{\binom{n-1}{m-1}}\sum_q \nu_{q;rs}\nu_{q;t,u}i^{tr}i^{us} + O(n^{-1}), \end{aligned} \tag{3.5}$$

where $\nu_{q;r,s} = E_{Y_q}\{\partial_r\ell(\omega^*;Y_q)\partial_s\ell(\omega^*;Y_q)\}$ and $\nu_{q;rs} = E_{Y_q}\{\partial_{rs}\ell(\omega^*;Y_q)\}$, for $r,s = 1,\ldots,d$, with $\partial_r\ell(\omega^*;Y_q) = \sum_{i\in q}\partial_r\ell(\omega^*;Y_i)$ and $\partial_{rs}\ell(\omega^*;Y_q) = \sum_{i\in q}\partial_{rs}\ell(\omega^*;Y_i)$.

The proof of Theorem 3.1 is deferred to the Appendix. For the case $\delta = 0$, namely when m is fixed, the following corollary holds. The proof is analogous to that of Theorem 3.1, with the additional simplification that the third term in the right hand side of equation (A.3), as given in the Appendix, is of order $O_p(n^{-1})$ and therefore included in the error term.

Corollary 3.1 *If m is fixed, the selection statistic (2.1) is such that*

$$E_Y\{\Psi_{CV(m)}(Y)\} = E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\,\nu_{t,r}i^{rt} - \frac{1}{\binom{n-1}{m-1}}\sum_q \nu_{q;r,s}i^{rs} + O(n^{-1}). \tag{3.6}$$

As a simple application of this last result, it is easy to see that the expansion for the case with $m = 1$, which corresponds to the simple cross-validation procedure, is

$$\begin{aligned} E_Y\{\Psi_{CV(1)}(Y)\} &= E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\,\nu_{t,r}i^{rt} - \sum_{i=1}^{n}\nu_{i;r,s}i^{rs} + O(n^{-1}) \\ &= E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\,\nu_{t,r}i^{rt} - \sum_{i=1}^{n}\nu_{i;r}\nu_{i;s}i^{rs} + O(n^{-1}). \end{aligned}$$

Here $\nu_{i;r,s} = E_{Y_i}\{\partial_r\ell(\omega^*;Y_i)\partial_s\ell(\omega^*;Y_i)\}$, $\nu_{i;r} = E_{Y_i}\{\partial_r\ell(\omega^*;Y_i)\}$, $r,s = 1,\ldots,d$, and it is easy to show that $\sum_{i=1}^n \nu_{i;r,s} = \nu_{r,s} + \sum_{i=1}^n \nu_{i;r}\nu_{i;s}$. Note that, in the particular situation with independent identically distributed observation, $\nu_{i;r} = 0$, $r = 1,\ldots,d$, exactly or to the relevant order of approximation. Thus, $\Psi_{CV(1)}(Y)$ is a first-order unbiased estimator for $\eta(g,\widehat{f})$ and $\nu_{t,r}i^{rt} = \bar{\nu}_{t,r}\bar{\imath}^{rt}$. Except this simple case, the first-order unbiasedness holds only if the model is true. Indeed, when the model is true, we can prove that $\Psi_{CV(1)}(Y) = \ell(\widehat{\omega};Y) - d + O_p(n^{-1/2})$, which corresponds to the AIC, in accordance with Stone's (1977a) result. The same conclusions are valid for the case with $m > 1$ fixed.

On the other hand, if m increases with n, so that $m = o(n)$, the expected behaviour of $\Psi_{CV(m)}(Y)$ differs. In particular, with independent identically distributed observations, we have that

$$\frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{q;r,s} i^{rs} = \frac{\binom{n}{m}}{\binom{n-1}{m-1}} \frac{m\, \bar{\nu}_{r,s} \bar{\imath}^{rs}}{n} = \bar{\nu}_{r,s} \bar{\imath}^{rs},$$

$$\frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{q;rs} \nu_{q;t,u} i^{tr} i^{us} = \frac{\binom{n}{m}}{\binom{n-1}{m-1}} \frac{m^2\, \bar{\nu}_{rs} \bar{\nu}_{t,u} \bar{\imath}^{tr} \bar{\imath}^{us}}{n^2} = -\frac{m}{n}\, \bar{\nu}_{t,r} \bar{\imath}^{rt},$$

where the expected likelihood quantity $\bar{\nu}_{rs}$ is related to a single component Y_i. Thus, (3.5) particularizes to

$$\begin{aligned} E_Y\{\Psi_{CV(m)}(Y)\} &= E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\, \bar{\nu}_{t,r} \bar{\imath}^{rt} - \bar{\nu}_{t,r} \bar{\imath}^{rt} - \frac{1}{2} \frac{m}{n}\, \bar{\nu}_{t,r} \bar{\imath}^{rt} + O(n^{-1}) \\ &= E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2} \left(1 + \frac{m}{n}\right) \bar{\nu}_{t,r} \bar{\imath}^{rt} + O(n^{-1}). \end{aligned} \tag{3.7}$$

Note that, in this case, $\Psi_{CV(m)}(Y)$ is an unbiased estimator for $\eta(g, \widehat{f})$ to order $o(1)$. The additional bias term $-m\, \bar{\nu}_{t,r} \bar{\imath}^{rt}/(2n)$ is of order $O(n^{\delta-1})$, so that it is negligible for large n. Moreover, if the model contains the true distribution, we have that $\bar{\nu}_{t,r} \bar{\imath}^{rt} = d$ and, from (A.3), we may prove that $\Psi_{CV(m)}(Y)$ is asymptotically equivalent to a suitable modification of the AIC. More precisely, if $\delta \in (0, 1/2]$ we find that $\Psi_{CV(m)}(Y) = \ell(\widehat{\omega}; Y) - d + O_p(n^{-1/2})$, whereas, if $\delta \in (1/2, 1)$, we have to consider an additional term of order $O(n^{\delta-1})$, so that $\Psi_{CV(m)}(Y) = \ell(\widehat{\omega}; Y) - d\{1 + m/(2n)\} + O_p(n^{-1/2})$.

Theorem 3.2 *Under the assumptions stated before, if $\delta \in (0, 1]$ and $\gamma \in (1/2, 1]$, the selection statistic (2.1) is such that*

$$E_Y\{\Psi_{CV(m)}(Y)\} = E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2} \frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{(q);t,u}\, i^{rt}_{(q)}\, i^{us}_{(q)}\, \nu_{q;rs} + O(n^{1-2\gamma}), \tag{3.8}$$

where $\nu_{(q);r,s} = E_{Y_{(q)}}\{\partial_r \ell(\omega^*; Y_{(q)}) \partial_s \ell(\omega^*; Y_{(q)})\}$, $\nu_{(q);rs} = E_{Y_{(q)}}\{\partial_{rs} \ell(\omega^*; Y_{(q)})\}$ and $i^{rs}_{(q)}$ is the (r, s)-element of the inverse of the matrix $[i_{(q);rs}] = [-\nu_{(q);rs}]$. Here, $\partial_r \ell(\omega^*; Y_{(q)}) = \sum_{i \notin q} \partial_r \ell(\omega^*; Y_i)$, $\partial_{rs} \ell(\omega^*; Y_{(q)}) = \sum_{i \notin q} \partial_{rs} \ell(\omega^*; Y_i)$ and the expectations are with respect to the true distribution of $Y_{(q)}$.

The proof of Theorem 3.2 is deferred to the Appendix. Note that Theorem 3.2 can not be applied when m is fixed or $\gamma \in (0, 1/2]$. However, it is useful for the case when $\delta = 1$, which is excluded in Theorem 3.1. In fact, the proof does not involve the asymptotic relationship between $\widehat{\omega}_{(q)}$ and $\widehat{\omega}$, which holds for $\delta \neq 1$, and it is similar to that considered for the approximation of the target value $\eta(g, \widehat{f})$, given in Proposition 3.1.

Finally, we emphasize that, with independent, identically distributed observations, we obtain

$$\frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{(q);t,u}\, \imath^{rt}_{(q)}\, \imath^{us}_{(q)}\, \nu_{q;rs} = \frac{\binom{n}{m}}{\binom{n-1}{m-1}} \frac{m(n-m)\, \bar{\nu}_{t,u} \bar{\nu}_{rs} \bar{\imath}^{rt} \bar{\imath}^{us}}{(n-m)^2} = -\frac{n}{(n-m)}\, \bar{\nu}_{t,r} \bar{\imath}^{rt}$$

and (3.8) particularizes to

$$E_Y\{\Psi_{CV(m)}(Y)\} = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2} \frac{n}{(n-m)}\, \bar{\nu}_{t,r} \bar{\imath}^{rt} + O(n^{1-2\gamma}). \tag{3.9}$$

Whenever $m = o(n)$, we have $\gamma = 1$ and $n/(n-m) = 1 + m/n + O(n^{2(\delta-1)})$, so that (3.9) corresponds to (3.7), as expected. By means of suitable expansions, we can prove that, whenever $\gamma \in (2/3, 1]$,

$$\Psi_{CV(m)}(Y) = \ell(\widehat{\omega}; Y) - \frac{1}{2} \left\{ 1 + \frac{n}{(n-m)} \right\} \bar{\nu}_{t,r} \bar{\imath}^{rt} + O_p(n^{1-3\gamma/2}). \tag{3.10}$$

Indeed, if the model contains the true distribution, we have that $\bar{\nu}_{t,r} \bar{\imath}^{rt} = d$ and (3.10) turns out to be a suitable modification of the AIC.

4 Comments

The theorems and the corollary, presented in the previous section, enable the calculation of the first-order bias term for both the simple and the multifold cross-validation estimators, under different assumptions on the model and on the dimension m of the validation set. These terms are obtained by simply comparing equations (3.5), (3.6) and (3.8) with the asymptotic approximation for the target value $\eta(g, \widehat{f})$ given by (3.2). These original results may be useful for understanding some theoretical aspects of the cross-validation procedures and for choosing an appropriate strategy, for a given model selection problem.

As a summary of the results presented in Section 3.2, we recall that the simple and the multifold cross-validation procedures, with m fixed, define a first-order unbiased estimator for $\eta(g, \widehat{f})$ in the particular case of independent, identically distributed, observations. However, this result holds in a more general context of non independent observations only if the candidate model is correctly specified. On the other hand, the multifold cross-validation procedure, when m increases with n, usually corresponds to a biased estimator for the expected predictive loglikelihood $\eta(g, \widehat{f})$ and the bias term turns out to be asymptotically negligible for large n provided that $m = o(n)$. An intuitive motivation, supporting this conclusion, concerns the fact that the second term in the right hand side of (3.8) is, in some sense, similar to the the second term in the right hand side of (3.2), which, recalling the proof of Proposition 3.1, corresponds to

$$\frac{1}{2}\, E_Y\{(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)\} E_Z\{\partial_{rs}\ell(\omega^*; Z)\} = \frac{1}{2}\, \nu_{t,u} i^{rt} i^{us} \nu_{rs} = -\frac{1}{2}\, \nu_{t,r} i^{rt}.$$

The term $\nu_{(q);t,u}\, i^{rt}_{(q)}\, i^{us}_{(q)}\, \nu_{q;rs}$ in equation (3.8) is similar to $\nu_{t,u} i^{rt} i^{us} \nu_{rs}$, since $Y_{(q)}$ and Y_q, for each subset $q \subseteq \{1, \ldots, n\}$ of size m, play the role of the observable random vector Y and future random vector Z, respectively. Thus, $\Psi_{CV(m)}(Y)$, with $m = O(n^\delta)$, $\delta \in (0, 1]$, aims to mimic $\eta(g, \widehat{f})$, where the dimension of both Y and Z increases with n.

As an additional comments on the results, we have to emphasized that we essentially obtain first-order approximations, where the remainder is usually expected to vanish as n increases. However, it may happen that the error term turns out to be be substantial even for a large dimension n of the observed sample. As a matter of fact, it may be heavily influenced by the dimension m of the validation set, by the underlying model, and in particular by the number of unknown parameters which could also increases with n, and by the fact that there might be a non-finite number of alternative models. Thus, if we go beyond a standard model selection framework, the behaviour of the cross-validation estimator may differ from that of its first-order approximation.

In this paper we focus on the potential bias of various cross-validation estimators, however, in order to properly evaluate their performances, a further study on the stability of the estimates should be required. A deep analysis, similar to that one performed by Burman (1989), giving the expansions for the variance of the v-fold cross-validation and the repeated learning-testing criteria, could be useful for classifying the cross-validation estimators according to the bias-variance trade-off criterion. Nevertheless this analysis is beyond the scope of the present paper, even if some preliminary hints on the variability issue may be deduced from the simulation study presented in Section 6.

A further comment concerns the computational cost of the cross-validation procedures which may become unbearable when n and eventually m are large. To simplify the calculations, we may consider equation (A.3), which is used in the proof of Theorem 3.1 presented in the Appendix. This relation holds for $\delta \in (0, 1)$, and, without the third term in the right hand side, for $\delta = 0$, that is with m fixed. From (A.3) we may obtain a first-order asymptotically equivalent expression for $\Psi_{CV(m)}(Y)$, based on the maximum likelihood estimator $\widehat{\omega}$ and defined as a modification of the maximized loglikelihood $\ell(\widehat{\omega}; Y)$. This can be useful for applications, since the computation of the selection statistic is in fact greatly simplified. These simplified formulas for the cross-validation estimators will be specified in Section 5 for the problem of variable selection in linear regression models.

A final interesting point to be discussed regards the usefulness of the above mentioned results

in the general context of model selection strategies. Although, as noted by Fushiki (2011), a better estimator for the prediction error, and in particular a bias-corrected version, does not necessarily correspond to a better model selection criterion, the findings on the first-order bias term of the cross-validation estimators may be useful as well in this more challenging context to understand the inferential properties of the estimators.

It is widely known that cross-validation is frequently considered for model selection, pointing to the model with the smallest estimated risk, however the validity of the selection procedure depends on the specific goal of the model selection. Under this respect, for density estimation purposes, Celisse (2014) shows that the performance of alternative cross-validation criteria varies according to the objective of the procedure, namely risk estimation or model selection in a strict sense. This distinction is emphasized in the review paper by Arlot and Celisse (2010), where, considering a more general framework, they speak about model selection for identification and model selection for estimation. In the first case, we assume that the true model exists and it belongs to the set of candidate model, and the aim is to identify it, or that the true model exists but it does not belong to the collection, and then we look for the candidate model closest to the true one. Here, the performance of the model selection procedures is usually evaluated in terms of model selection consistency. In the second case, we do not necessarily assume the existence of a true model and the aim is to find the model which minimizes a suitable estimated risk. The performance of the model selection procedures is then evaluated in terms of model selection efficiency, with respect to an ideal model which minimizes the true risk. With regard to the regression framework, an interesting contribution on the conflict between model identification and model estimation is Yang (2005), while Yang (2007) presents a deep analysis on the conditions assuring model selection consistency for cross-validation selection criteria.

Since we assume the existence of a true model, belonging or not to the collection of candidate models, the cross-validation estimators studied in present paper can be ideally employed for defining model selection criteria designed for the goal of model identification, where the best candidate model is specified as that one minimizing the estimated Kullback-Liebler risk. However, the original contribution of this paper does not regard the optimality issues concerning model selection, since the focus is mainly on first-order bias evaluation, which could be a preliminary task for understanding the behaviour of the alternative selection criteria. In the simulation study of Section 6, we presents a preliminary analysis of how the cross-validation estimators, used as model selection statistics in the variable selection framework, behave with respect to different choices for m, obtaining results in

accordance with the theoretical findings of Shao (1997) and Yang (2007).

5 Applications

5.1 Normal regression models

Let us assume that the candidate model for the observations $y = (y_1, \ldots, y_n)^T$, $n \geq 1$, is specified by the family of joint density functions $f(y;\omega)$, $\omega \in \Omega \subseteq \mathbf{R}^d$, $d > 1$, for a random vector $Y = (Y_1, \ldots, Y_n)^T$ such that

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

with $\mu_i = \eta^{-1}(\beta^T x_i)$. Here, $\beta = (\beta_1, \ldots, \beta_{d-1})^T$ is a $(d-1)$-dimensional vector of unknown parameters, $x_i = (x_{i1}, \ldots, x_{id-1})^T$ is a $(d-1)$-dimensional vector of known covariates and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ follows a multivariate normal distribution $N_n(\mathbf{0}, \sigma^2\mathbf{I})$, with $\mathbf{0}$ the null vector, $\mathbf{I}$ the identity matrix and $\sigma^2 > 0$ an unknown parameter. Thus, $Y_1, \ldots, Y_n$ are independent Gaussian random variables such that Y_i follows a $N(\mu_i, \sigma^2)$ distribution, $i = 1, \ldots, n$. Indeed, $\eta(\cdot)$ is a suitable monotonic differentiable link function with inverse $\eta^{-1}(\cdot)$. The simplified model with $\mu_i = \eta^{-1}(\beta^T x_i) = \beta^T x_i$ is obtained whenever the canonical link function $\eta(u) = u$ is considered and corresponds to the linear Gaussian regression model. Moreover, the unknown parameter is $\omega = (\omega_1, \ldots, \omega_d)^T = (\beta_1, \ldots, \beta_{d-1}, \sigma^2)^T = (\beta^T, \sigma^2)^T$ and, whenever $g(y) = f(y;\omega_0)$, $\omega_0 \in int(\Omega)$, that is, when the model contains the true distribution, ω^* equals the true parameter value $\omega_0 = (\beta_0^T, \sigma_0^2)^T$.

In this context, the assumptions introduced in Section 2.2 are fulfilled. Indeed, we introduce the following additional assumptions regarding the covariates: the $n \times (d-1)$-dimensional matrix $X = [x_{ir}]$, $i = 1, \ldots, n$, $r = 1, \ldots, d-1$, is such that

$$(X^T X)^{-1} = \Sigma\, n^{-1} + O(n^{-2}),$$

with $\Sigma = [\Sigma_{rs}]$, $r, s = 1, \ldots, d-1$, a known $(d-1) \times (d-1)$-dimensional matrix; the matrices $X_{(q)} = [x_{ir}]$, $i \notin q$, $r = 1, \ldots, d-1$, and $X_q = [x_{ir}]$, $i \in q$, $r = 1, \ldots, d-1$, are such that

$$(X_{(q)}^T X_{(q)})^{-1} = \Sigma\,(n-m)^{-1} + O((n-m)^{-2}), \quad (X_q^T X_q)^{-1} = \Sigma\, m^{-1} + O(m^{-2}),$$

whenever $n - m$ and m increase with n.

The maximum likelihood estimators for ω correspond to $\widehat{\omega} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{d-1}, \widehat{\sigma}^2)^T = (\widehat{\beta}^T, \widehat{\sigma}^2)^T$ and $\widehat{\omega}_{(q)} = (\widehat{\beta}_{(q)1}, \ldots, \widehat{\beta}_{(q)d-1}, \widehat{\sigma}^2_{(q)})^T = (\widehat{\beta}^T_{(q)}, \widehat{\sigma}^2_{(q)})^T$, defined as the solutions, with respect to ω, to (2.2) and (2.3), respectively, with

$$\partial_r \ell(\omega; Y_i) = \frac{(Y_i - \mu_i) x_{ir}}{\sigma^2 \eta'(\mu_i)}, \quad r = 1, \ldots, d-1,$$

$$\partial_r \ell(\omega; Y_i) = \partial_{\sigma^2} \ell(\omega; Y_i) = \frac{1}{2\sigma^4} \{(Y_i - \mu_i)^2 - \sigma^2\}, \quad r = d,$$

where $\eta'(\cdot)$ is the first derivative of function $\eta(\cdot)$. In the canonical case we have that $\eta'(\mu_i) = 1$. Note that $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2$ and $\widehat{\sigma}^2_{(q)} = (n-m)^{-1} \sum_{j \notin q} (Y_j - \widehat{\mu}_{(q)j})^2$, with $\widehat{\mu}_i = \eta^{-1}(\widehat{\beta}^T x_i)$ and $\widehat{\mu}_{(q)j} = \eta^{-1}(\widehat{\beta}^T_{(q)} x_j)$; in particular, if $q = \{i\}$, $\widehat{\sigma}^2_{(q)} = \widehat{\sigma}^2_{(i)} = (n-1)^{-1} \sum_{j \neq i} (Y_j - \widehat{\mu}_{(i)j})^2$, with $\widehat{\mu}_{(i)j} = \eta^{-1}(\widehat{\beta}^T_{(i)} x_j)$.

In this framework, the cross-validation selection statistic corresponds to

$$\Psi_{CV(m)}(Y) = -\frac{n}{2} \log(2\pi) - \frac{m}{2\binom{n-1}{m-1}} \sum_q \log \widehat{\sigma}^2_{(q)} - \frac{1}{2\binom{n-1}{m-1}} \sum_q \sum_{i \in q} \frac{(Y_i - \widehat{\mu}_{(q)i})^2}{\widehat{\sigma}^2_{(q)}},$$

and, for the leave-one-out case, it simplifies to

$$\Psi_{CV(1)}(Y) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \log \widehat{\sigma}^2_{(i)} - \frac{1}{2} \sum_{i=1}^n \frac{(Y_i - \widehat{\mu}_{(i)i})^2}{\widehat{\sigma}^2_{(i)}}.$$

These formulas require the computation of the maximum likelihood estimator $\widehat{\omega}_{(q)}$ for all the subsets $Y_{(q)}$ of vector Y. In order to reduce the computations, we may use the relation linking $\widehat{\omega}_{(q)}$ and $\widehat{\omega}$, derived in the following corollary of Proposition 2.1.

Corollary 5.1 *Under the assumptions stated in Section 2.2 with $\delta \neq 1$, if we consider a normal regression model with link function $\eta(\cdot)$, we have that*

$$\widehat{\beta}_{(q)u} = \widehat{\beta}_u + \sum_{j \in q} \frac{(Y_j - \widehat{\mu}_j) x_{jr}}{\widehat{\sigma}^2 \eta'(\widehat{\mu}_j)} \, \partial^{ru} \ell(\widehat{\omega}; Y) + o_p(n^{\delta - 1}), \quad u = 1, \ldots, d-1,$$

$$\widehat{\sigma}^2_{(q)} = \widehat{\sigma}^2 + \frac{m}{n} \widehat{\sigma}^2 - \frac{1}{n} \sum_{j \in q} (Y_j - \widehat{\mu}_j)^2 + o_p(n^{\delta - 1}),$$

where $\partial^{ru} \ell(\widehat{\omega}; Y)$ is the (r,u)-element of the inverse of matrix $[\partial_{ru} \ell(\widehat{\omega}; Y)]$, $r, u = 1, \ldots, d-1$, defined by (A.5).

The proof of Corollary 5.1 is deferred to the Appendix. It may be useful to emphasize that these relations hold for $\delta \neq 1$, that is when m is fixed or $m = o(n)$. The formulas for $\widehat{\beta}_{(i)}$ and $\widehat{\sigma}^2_{(i)}$, useful within the simple cross-validation procedure, may be obtained by setting $m = 1$ and $q = \{i\}$. Note that, if the canonical link function is considered, the expansion for $\widehat{\beta}_{(q)u}$ simplifies, since $\partial_{ru}\ell(\widehat{\omega}; Y) = -\widehat{\sigma}^{-2}\sum_{i=1}^n x_{ir}x_{iu}$ and $\partial^{ru}\ell(\widehat{\omega}; Y) = -\widehat{\sigma}^2\Sigma_{ru}n^{-1} + O(n^{-2})$, $r, u = 1, \ldots, d-1$, so that

$$\widehat{\beta}_{(q)u} = \widehat{\beta}_u - \frac{1}{n}\sum_{j\in q}(Y_j - \widehat{\mu}_j)x_{jr}\Sigma^{ru} + O_p(n^{-2}), \quad u = 1, \ldots, d-1.$$

In order to speed up the computation of $\Psi_{CV(m)}(Y)$, we may also consider some asymptotically equivalent expressions based on the maximum likelihood estimator $\widehat{\omega}$ and derived from relation (A.3), obtained in the proof of Theorem 3.1. Whenever m is fixed, the third term in the right hand side of (A.3) has to be included in the error term and we get

$$\begin{aligned}
\Psi_{CV(m)}(Y) &= -\frac{n}{2}\left\{\log(2\pi\widehat{\sigma}^2) + 1\right\} - \frac{1}{2\widehat{\sigma}^4 n\binom{n-1}{m-1}}\sum_q\sum_{i\in q}(Y_i - \widehat{\mu}_i)^4 + \frac{1}{\widehat{\sigma}^2 n\binom{n-1}{m-1}}\sum_q\sum_{i\in q}(Y_i - \widehat{\mu}_i)^2 \\
&\quad -\frac{1}{2} + \frac{1}{\widehat{\sigma}^4\binom{n-1}{m-1}}\sum_q\left\{\sum_{i\in q}\frac{(Y_i - \widehat{\mu}_i)x_{ir}}{\eta'(\widehat{\mu}_i)}\sum_{j\in q}\frac{(Y_j - \widehat{\mu}_j)x_{js}}{\eta'(\widehat{\mu}_j)}\right\}\partial^{rs}\ell(\widehat{\omega}; Y) + O_p(n^{-1}) \\
&= -\frac{n}{2}\left\{\log(2\pi\widehat{\sigma}^2) + 1\right\} - \frac{1}{2\widehat{\sigma}^4 n}\sum_{i=1}^n(Y_i - \widehat{\mu}_i)^4 + \frac{1}{2} \\
&\quad + \frac{1}{\widehat{\sigma}^4\binom{n-1}{m-1}}\sum_q\left\{\sum_{i\in q}\frac{(Y_i - \widehat{\mu}_i)x_{ir}}{\eta'(\widehat{\mu}_i)}\sum_{j\in q}\frac{(Y_j - \widehat{\mu}_j)x_{js}}{\eta'(\widehat{\mu}_j)}\right\}\partial^{rs}\ell(\widehat{\omega}; Y) + O_p(n^{-1}).
\end{aligned}$$

This final expression is obtained by noticing that $\sum_q\sum_{i\in q}(Y_i - \widehat{\mu}_i)^\tau = \binom{n-1}{m-1}\sum_{i=1}^n(Y_i - \widehat{\mu}_i)^\tau$, with $\tau = 2, 4$. In particular, if $m = 1$ and $q = \{i\}$ we have that

$$\begin{aligned}
\Psi_{CV(1)}(Y) &= -\frac{n}{2}\left\{\log(2\pi\widehat{\sigma}^2) + 1\right\} - \frac{1}{2\widehat{\sigma}^4 n}\sum_{i=1}^n(Y_i - \widehat{\mu}_i)^4 + \frac{1}{2} \\
&\quad + \frac{1}{\widehat{\sigma}^4}\sum_{i=1}^n\left[\frac{(Y_i - \widehat{\mu}_i)^2 x_{ir}x_{is}}{\{\eta'(\widehat{\mu}_i)\}^2}\right]\partial^{rs}\ell(\widehat{\omega}; Y) + O_p(n^{-1}).
\end{aligned}$$

Moreover, if $m = o(n)$, we may derive a suitable asymptotic expansion for $\Psi_{CV(m)}(Y)$, which corresponds to that one given for m fixed, with an additional term based on $\partial_r\ell(\widehat{\omega}; Y_q)$, $\partial_{rs}\ell(\widehat{\omega}; Y_q)$, $\partial^{rs}\ell(\widehat{\omega}; Y)$. In the canonical case, all these formulas are in fact simplified. Finally, if $m = O(n)$ analogous results are not available.

Concerning the evaluation of the potential bias of $\Psi_{CV(m)}(Y)$, we need to compute the expectation of the selection statistic under the true distribution, approximated by means of the asymptotic relations given in Section 3. These approximations require the calculation of some expected likelihood quantities, with respect to the true distribution of Y, related to the candidate model and to the pseudo-true parameter value ω^*. Unless a suitable assumption on the true distribution is considered, these quantities are not explicitly known and they may be eventually estimated by using non-parametric bootstrap techniques. In the following, we shall derive explicitly these expressions for the special case of a linear Gaussian regression model.

5.2 Variable selection in linear regression models

In this section, we focus on the problem of variable selection under a linear Gaussian regression model, namely a Gaussian regression model with a canonical link function. In particular, we assume that the true model has $d_0 - 2$, with $d_0 \geq 2$, covariates; namely, $Y = (Y_1, \ldots, Y_n)^T$ is such that

$$Y_i = \mu_{0i} + \varepsilon_{0i}, \quad i = 1, \ldots, n,$$

where $\mu_{0i} = \beta_0^T x_{0i}$, with $\beta_0 = (\beta_{01}, \ldots, \beta_{0d_0-1})^T$, $x_{0i} = (x_{i1}, \ldots, x_{id_0-1})^T$ and $\varepsilon_0 = (\varepsilon_{01}, \ldots, \varepsilon_{0n})^T$ follows a multivariate normal distribution $N_n(\mathbf{0}, \sigma_0^2 \mathbf{I})$, with $\sigma_0^2 > 0$. Let us consider the candidate model as defined in Section 5.1, with $\eta(u) = u$. Two different situations have to be considered. In the first case, $d \geq d_0$ so that the model is correct. However, unless we have $d = d_0$, the model may be inefficient, because of its unnecessarily large size. In the second case, $d < d_0$ so that the model is incorrect, since some relevant covariates are not taken into account.

Let us consider the first situation, where the model is correct but it may include some redundant terms in the linear predictor $\beta^T x_i$. In this case, the true parameter value β_0 may be specified as the $(d-1)$-dimensional vector $\beta_0 = (\beta_{01}, \ldots, \beta_{0d_0-1}, 0, \ldots, 0)^T$, with, eventually, $d - d_0$ null components; indeed, $\omega^* = \omega_0 = (\beta_0^T, \sigma_0^2)^T$. The following proposition provides the first-order asymptotic expansions for the target quantity $\eta(g, \widehat{f})$ and for the expectation $E_Y\{\Psi_{CV(m)}(Y)\}$, by considering alternative asymptotic regimes.

Proposition 5.1 *Under the assumptions recalled in Section 5.1, with $m = O(n^\delta)$ and $n - m = O(n^\gamma)$, if we consider a linear Gaussian regression model with $d \geq d_0$, we have that*

$$\eta(g, \widehat{f}) = -\frac{n}{2}\left\{\log(2\pi\sigma_0^2) + 1\right\} - \frac{d}{2} + O(n^{-1}), \tag{5.1}$$

$$E_Y\{\Psi_{CV(m)}(Y)\} = -\frac{n}{2}\left\{\log(2\pi\sigma_0^2)+1\right\} - \frac{d}{2}\left(1+\frac{m}{n}\right) + O(n^{-1}), \tag{5.2}$$

for $\delta \in (0,1)$, $\gamma \in (0,1]$,

$$E_Y\{\Psi_{CV(m)}(Y)\} = -\frac{n}{2}\left\{\log(2\pi\sigma_0^2)+1\right\} - \frac{d}{2}\frac{n}{(n-m)} + O(n^{1-2\gamma}), \tag{5.3}$$

for $\delta = 1$, $\gamma \in (1/2,1]$.

The proof of Proposition 5.1 is deferred to the Appendix. It is easy to see that, whenever m is fixed and, in particular, for the leave-one-out case, we have $\delta = 0$ and $\gamma = 1$ and the expectation is simply obtained by (5.2), neglecting the term m/n. Moreover, in the special case when $m/n = \lambda + o(1)$, with $\lambda \in (0,1)$, we may substitute $1/(1-\lambda) + o(1)$ for $n/(n-m)$ in equation (5.3). Thus, as far as linear Gaussian regression models are concerned, the simple cross-validation selection statistic and, in general, the multifold cross-validation selection statistics with m fixed, are first order unbiased estimators for $\eta(g,\widehat{f})$. However, this conclusion requires the key, but rarely plausible, assumption that the model is correctly specified. On the other hand, if m increases with n, $\Psi_{CV(m)}(Y)$ is a biased estimator for $\eta(g,\widehat{f})$, unless $\gamma = 1$ and $\delta \in (0,1)$, so that $m = o(n)$.

Finally, we can prove that the same asymptotic relations linking the cross-validation selection statistics and the AIC, presented in Section 3 for the the case of independent, identically distributed, observations, maintain as well when we consider a (correctly specified) linear Gaussian regression model. In particular, we find that, if $m/n = \lambda + o(1)$, as $n \to +\infty$, with $\lambda \in (0,1)$, since $\delta = 1$ and $\gamma = 1$, $\Psi_{CV(m)}(Y) = \ell(\widehat{\omega};Y) - d(2-\lambda)/\{2(1-\lambda)\} + O_p(n^{-1/2})$. This result is in accordance with that one given by Zhang (1993), with regard to the problem of selecting among alternative linear regression models, whenever the model is the correct one and σ^2 is supposed to be known.

Let us consider the second situation where the model is not correctly specified, since some relevant covariates are lacking. In this case, $d < d_0$ and the true parameter values are σ_0^2 and $\beta_0 = (\beta_{A0}^T, \beta_{B0}^T)^T = (\beta_{01},\ldots,\beta_{0d_0-1})^T$, with $\beta_{A0} = (\beta_{01},\ldots,\beta_{0d-1})^T$, $\beta_{B0} = (\beta_{0d},\ldots,\beta_{0d_0-1})^T$, while $\beta = (\beta_1,\ldots,\beta_{d-1})^T$ is the $(d-1)$-dimensional parameter specified by the assumed model. The vectors including all the relevant covariates are $x_{0i} = (x_{Ai}^T, x_{Bi}^T)^T = (x_{i1},\ldots,x_{id_0-1})^T$, $i = 1,\ldots,n$, with $x_{Ai} = x_i = (x_{i1},\ldots,x_{id-1})^T$ the covariate vector defining the model under consideration and $x_{Bi} = (x_{id},\ldots,x_{id_0-1})^T$. We state a further assumption on the covariates, assuring that the $n\times(d-1)$-dimensional matrix $X = X_A = [x_{ir}]$, $i = 1,\ldots,n$, $r = 1,\ldots,d-1$, and the $n \times (d_0-d)$-dimensional matrix $X_B = [x_{ir}]$, $i = 1,\ldots,n$, $r = d,\ldots,d_0-1$, are such that

$$X^T X_B = X_A^T X_B = n\,\Gamma + O(1), \quad X_B^T X_B = n\,\Delta + O(1),$$

with $\Gamma = [\Gamma_{rs}]$, $r = 1, \ldots, d-1$, $s = 1, \ldots, d_0 - d$, a known $(d-1) \times (d_0 - d)$-dimensional matrix and $\Delta = [\Delta_{tu}]$, $t, u = 1, \ldots, d_0 - d$, a known $(d_0 - d) \times (d_0 - d)$-dimensional matrix. These additional requirements on the covariates are rather mild and, together with the assumptions made on X at the beginning of Section 5.1, assure that the usual asymptotic results hold whenever the model is misspecified.

Moreover, by solving equation (2.4) with respect to $\omega = (\beta^T, \sigma^2)^T$, we find that $\omega_n^* = (\beta_n^{*T}, \sigma_n^{*2})^T$, $n \geq 1$, and $\omega^* = (\beta^{*T}, \sigma^{*2})^T$ are such that $\omega_n^* = \omega^* + O(n^{-1})$, with

$$\beta^* = \beta_{A0} + \Sigma\,\Gamma\beta_{B0}, \quad \sigma^{*2} = \sigma_0^2 + \beta_{B0}^T(\Delta - \Gamma^T\Sigma\,\Gamma)\beta_{B0} + O(n^{-1}) = \sigma_0^2(1 + \rho/n) + O(n^{-1}).$$

Here,

$$\rho = \sigma_0^{-2}\sum_{i=1}^{n}(\beta_0^T x_{0i} - \beta_n^{*T} x_i)^2 = \sigma_0^{-2}\{\beta_{B0}^T(\Delta - \Gamma^T\Sigma\,\Gamma)\beta_{B0}\}n + O(1)$$

is the non centrality parameter for the chi-squared distributed random variable $n\widehat{\sigma}^2/\sigma_0^2$, with $n-d+1$ degrees of freedom.

The following proposition provides the first-order asymptotic expansions for the target quantity $\eta(g, \widehat{f})$ and for the expectation $E_Y\{\Psi_{CV(m)}(Y)\}$, assuming m fixed and m increasing with n.

Proposition 5.2 *Under the assumptions stated before, with $m = O(n^\delta)$ and $n - m = O(n^\gamma)$, if we consider a linear Gaussian regression model with $d < d_0$, we have that*

$$\eta(g, \widehat{f}) = E_Y\{\log f(Y; \omega^*)\} - \frac{1}{2}\left\{\frac{d}{1 + \rho/n} + \frac{\rho/n}{(1 + \rho/n)^2}\right\} + O(n^{-1}), \tag{5.4}$$

$$E_Y\{\Psi_{CV(m)}(Y)\} = \eta(g, \widehat{f}) - (d-1)\frac{\rho/n}{1 + \rho/n} + O(n^{-1}), \tag{5.5}$$

for $\delta = 0$, $\gamma = 1$ (m fixed),

$$E_Y\{\Psi_{CV(m)}(Y)\} = \eta(g, \widehat{f}) - \frac{1}{2}\frac{m}{(n-m)}\left\{\frac{d}{1 + \rho/n} + \frac{\rho/n}{(1 + \rho/n)^2}\right\} + O(n^{1-2\gamma}), \tag{5.6}$$

for $\delta \in (0, 1]$, $\gamma \in (1/2, 1]$.

The proof of Proposition 5.2 is deferred to the Appendix. It is almost immediate to see that when $d = d_0$, that is when the model is correct, we have $\rho = 0$ and the above results equals those ones obtained in Proposition 5.1.

Whenever m is fixed, since $\rho = O(n)$, $\Psi_{CV(m)}(Y)$ could be a downwardly biased estimator for the expected predictive loglikelihood. Note that, to the relevant order of approximation, this result does not depend on m, so that the same conclusion is valid for $m = 1$, namely for the simple cross-validation selection statistic $\Psi_{CV(1)}(Y)$. With regard to the AIC, it is easy to see that

$$\begin{aligned} E_Y\{\ell(\widehat{\omega}; Y) - d\} &= E_Y[\log f(Y; \omega^*)] + \frac{1}{2}\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} - d + O(n^{-1}) \\ &= \eta(g, \widehat{f}) - d\frac{\rho/n}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2} + O(n^{-1}), \end{aligned}$$

so that it turns out to be even more biased than $\Psi_{CV(m)}(Y)$.

Furthermore, for the case where $\delta \in (0, 1]$, $\gamma \in (1/2, 1]$, $\Psi_{CV(m)}(Y)$ is still a biased estimator for $\eta(g, \widehat{f})$. However, if $\gamma = 1$ and $\delta \in (0, 1)$, so that $m = o(n)$, the first-order bias term is asymptotically negligible. In the special case when $m/n = \lambda + o(1)$, $\lambda \in (0, 1)$, we may substitute $\lambda/(1-\lambda) + o(1)$ for $m/(n-m)$. In this situation, if $\lambda \to 0$, we obtain a first order unbiased estimator for the target mean value.

Finally, as for the case with a correctly specified model, we may derive asymptotically equivalent expressions for $\Psi_{CV(m)}(Y)$, defined as a suitable modifications of the maximized loglikelihood $\ell(\widehat{\omega}; Y)$. More precisely, if m is fixed, we obtain that

$$\begin{aligned} \Psi_{CV(m)}(Y) &= \ell(\widehat{\omega}; Y) - \left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} - (d-1)\frac{\rho/n}{1+\rho/n} + O_p(n^{-1/2}) \\ &= \ell(\widehat{\omega}; Y) - \left\{\frac{d + (d-1)\rho/n}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O_p(n^{-1/2}). \end{aligned}$$

Indeed, if m increases with n, assuming $\delta \in (0, 1]$ and $\gamma \in (1/2, 1]$, we get

$$\begin{aligned} \Psi_{CV(m)}(Y) &= \ell(\widehat{\omega}; Y) - \frac{1}{2}\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} \\ &\quad - \frac{1}{2}\frac{n}{(n-m)}\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O(n^{1-3\gamma/2}) \\ &= \ell(\widehat{\omega}; Y) - \frac{1}{2}\left(1 + \frac{n}{n-m}\right)\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O_p(n^{1-3\gamma/2}). \end{aligned} \tag{5.7}$$

In particular, when $m/n = \lambda + o(1)$, $\lambda \in (0, 1)$, we have $\delta = \gamma = 1$ and the term $1 + n/(n-m)$ can be substituted by $(2-\lambda)/(1-\lambda)$. In both cases, $\Psi_{CV(m)}(Y)$ may be viewed as a suitable modification of the AIC. Note that, if $d = d_0$, we have $\rho = 0$ and all these results equal those ones obtained in the case where the model is correct.

6 A simulation study

We close the paper with a simple simulation study based on the STEAM data provided by Draper and Smith (1998, Appendix 1). The response variable corresponds to the pounds of steam used monthly, with regard to the steam plant which is part of an industry; there are 9 explanatory variables and the number of observations is 25. The observed design matrix is extended in order to account for at most $n = 200$ observations; the additional 175 values for each covariate are obtained by random sampling of the 25 observed values.

We consider the following two situations. In the first case, the true model is the linear Gaussian regression model with $d_0 = 6$ parameters, namely $\beta_{01}, \ldots, \beta_{05}, \sigma_0^2$, which equal the fitted parameter values obtained from the STEAM data. Indeed, the candidate model is correct but it includes 3 further redundant covariates, so that $d = 9$. In the second case, the true model is the linear Gaussian regression model with $d_0 = 11$ parameters, namely $\beta_{01}, \ldots, \beta_{010}, \sigma_0^2$, which equal the fitted parameter values obtained from the complete STEAM dataset. Now the candidate model is incorrect since it includes only the first 5 covariates, so that $d = 7$. In both cases, β_{01} is the intercept parameter.

The aim of the study is to estimate the bias and the standard deviation of $\Psi_{CV(m)}(Y)$, interpreted as a suitable estimator for the target quantity $\eta(g, \widehat{f})$, with $n = 25, 50, 100$ and $m = 1, 3, 5, 10, 20$. The estimates for $\eta(g, \widehat{f})$ are obtained using a double Monte Carlo procedure based on 500×500 n-dimensional simulated samples from the true models for Y and Z, respectively. The mean and the standard deviation of $\Psi_{CV(m)}(Y)$ are estimated by considering $1,000$ n-dimensional simulated samples from the true model for Y. Whenever $m > 1$, for reducing the amount of computation involved, we consider the approximation to $\Psi_{CV(m)}(Y)$ proposed by Zhang (1993, Equation 1.3).

The estimated values are presented in Table 1, for the case where the model is true, and in Table 2, for the situation where the model is not correct. It is immediate to see that, in both situations, the case where $n = 25$ presents unstable results and the estimates for $n = 25$ and $m = 20$ are, as expected, not available. Indeed, it is interesting to note that, for m fixed, the bias of $\Psi_{CV(m)}(Y)$ tends to vanish as n increases only when the candidate model is correct (Table 1). On the other hand, when the model is not correctly specified (Table 2), $\Psi_{CV(m)}(Y)$ seems to be a biased estimator for the expected predictive loglikelihood $\eta(g, \widehat{f})$. This is in accordance with the theoretical findings presented in Section 5, regarding incorrect candidate models and m fixed. Finally, the standard deviation of $\Psi_{CV(m)}(Y)$, which is also influenced by the dimension of the target quantity $\eta(g, \widehat{f})$, is always higher when the model is correct and overparametrized. We presume that this phenomenon is related to the

n	25		50		100	
$\eta(g, \widehat{f})$	-51.969		-90.883		-175.061	
	Bias	SD	Bias	SD	Bias	SD
m						
1	-6,270	8,511	-1,839	6,025	-0,554	7,038
3	-10,898	9,531	-2,484	6,238	-1,115	7,533
5	-17,243	11,335	-3,070	6,238	-0,559	7,398
10	-70,561	20,350	-5,568	5,948	-1,428	7,662
20			-14,487	7,112	-2,303	7,084

Table 1: Monte Carlo estimates for the bias and the standard deviation of $\Psi_{CV(m)}(Y)$, viewed as estimator for $\eta(g, \widehat{f})$, with $n = 25, 50, 100$ and $m = 1, 3, 5, 10, 20$. The target $\eta(g, \widehat{f})$ is estimated by a double Monte Carlo procedure. Simulations of n-dimensional samples from the true linear Gaussian regression model with $d_0 = 6$ parameters estimated from the STEAM dataset. The candidate model is correct but it includes 3 unnecessary covariates.

n	25		50		100	
$\eta(g, \widehat{f})$	-39.917		-98.644		-211.399	
	Bias	SD	Bias	SD	Bias	SD
m						
1	-10,876	3,587	-7,919	2,428	-7,222	2,733
3	-14,043	4,118	-8,264	2,380	-7,313	2,838
5	-18,683	4,562	-8,712	2,497	-7,391	2,847
10	-43,519	8,664	-10,321	2,333	-7,793	2,841
20			-16,103	2,673	-8,664	2,950

Table 2: Monte Carlo estimates for the bias and the standard deviation of $\Psi_{CV(m)}(Y)$, viewed as estimator for $\eta(g, \widehat{f})$, with $n = 25, 50, 100$ and $m = 1, 3, 5, 10, 20$. The target $\eta(g, \widehat{f})$ is estimated by a double Monte Carlo procedure. Simulations of n-dimensional samples from the true linear Gaussian regression model with $d_0 = 11$ parameters estimated from the complete STEAM dataset. The candidate model is incorrect since it includes only the first 5 covariates.

fact that overfitting usually produces unstable predictive results.

In order to analyze the usefulness of the alternative cross-validation procedures for the specific aim of model identification, we perform an additional simulation experiment to evaluate the probability of correct selection, with alternative choices for the dimension m of the validation set. In this case we assume that the true model is that one with the first 5 covariates, which corresponds to the

linear Gaussian regression model with $d_0 = 7$ parameters, namely $\beta_{01}, \ldots, \beta_{06}, \sigma_0^2$, set equal the fitted parameter values obtained from the original dataset. The candidate regression models are nested and they have covariates ranging from 1 to 9, according to the order specified in the STEAM data.

The aim here is to estimate the probability of selecting the 9 candidate models using the selection statistic $\Psi_{CV(m)}(Y)$, with $n = 50, 100, 200$ and $m = 1, 5, 10, 20, 30, 50, 70, 100, 150$. The estimates are obtained by considering $1,000$ n-dimensional simulated samples from the true model. Instead of calculating $\Psi_{CV(m)}(Y)$ using the definition (2.1), which could be computationally demanding, we employ the first-order approximation defined by (5.7).

The estimated values are presented in Tables 3-5, for $n = 50, 100, 200$, respectively. We do not report the values for m nearly equal to n, since as expected we obtain unstable results. It is immediate to note that the probability of choosing the correct model is always the higher, whatever the value of m. However, even if the probability of correct selection increases with n, it is quite clear that, for a fixed n, the chance of choosing the correct model greatly increases with m. These findings, even though related to the particular case of variable selection in linear regression models, confirm the theoretical results of Shao (1997) and Yang (2007). In particular, they find that, at least for linear model selection, multifold cross-validation procedures are asymptotically consistent whenever $m/n \to 1$, as n increases. Moreover, this fact support the statement that the usefulness of a selection criterion depends on the objective of the model selection procedure, namely estimation or identification as recalled in Section 4, and that an unbiased selection statistic does not necessarily define an optimal criterion for model identification.

		$d-2$								
n	m	1	2	3	4	5	6	7	8	9
50	1	0.000	0.002	0.029	0.016	0.714	0.117	0.067	0.039	0.016
	5	0.000	0.004	0.031	0.016	0.730	0.106	0.064	0.035	0.014
	10	0.000	0.005	0.040	0.017	0.752	0.095	0.053	0.028	0.010
	20	0.000	0.008	0.085	0.025	0.766	0.069	0.029	0.014	0.004
	30	0.000	0.014	0.186	0.023	0.732	0.032	0.010	0.003	0.000

Table 3: Monte Carlo estimates for the probability of selecting alternative models using $\Psi_{CV(m)}(Y)$, with $n = 50$ and $m = 1, 5, 10, 20, 30$. Estimates based on n-dimensional simulated samples from the true linear Gaussian regression model with 5 covariates. The candidate regression models are nested and they have a number $d-2$ of covariates ranging from 1 to 9.

		$d-2$								
n	m	1	2	3	4	5	6	7	8	9
100	1	0.000	0.000	0.000	0.000	0.765	0.119	0.051	0.035	0.030
	5	0.000	0.000	0.000	0.000	0.777	0.113	0.049	0.033	0.028
	10	0.000	0.000	0.000	0.000	0.787	0.115	0.047	0.029	0.022
	20	0.000	0.000	0.000	0.000	0.810	0.109	0.040	0.025	0.016
	30	0.000	0.000	0.000	0.000	0.839	0.097	0.034	0.018	0.012
	50	0.000	0.000	0.000	0.000	0.904	0.069	0.021	0.004	0.002
	70	0.000	0.000	0.000	0.000	0.977	0.022	0.001	0.000	0.000

Table 4: Monte Carlo estimates for the probability of selecting alternative models using $\Psi_{CV(m)}(Y)$, with $n = 100$ and $m = 1, 5, 10, 20, 30, 50, 70$. Estimates based on n-dimensional simulated samples from the true linear Gaussian regression model with 5 covariates. The candidate regression models are nested and they have a number $d-2$ of covariates ranging from 1 to 9.

		$d-2$								
n	m	1	2	3	4	5	6	7	8	9
200	1	0.000	0.000	0.000	0.000	0.757	0.109	0.052	0.047	0.035
	5	0.000	0.000	0.000	0.000	0.759	0.109	0.051	0.046	0.035
	10	0.000	0.000	0.000	0.000	0.763	0.110	0.048	0.045	0.034
	20	0.000	0.000	0.000	0.000	0.774	0.107	0.045	0.043	0.031
	30	0.000	0.000	0.000	0.000	0.787	0.104	0.042	0.040	0.027
	50	0.000	0.000	0.000	0.000	0.819	0.102	0.031	0.032	0.016
	70	0.000	0.000	0.000	0.000	0.850	0.088	0.027	0.020	0.015
	100	0.000	0.000	0.000	0.000	0.895	0.069	0.021	0.008	0.007
	150	0.000	0.000	0.000	0.000	0.978	0.020	0.002	0.000	0.000

Table 5: Monte Carlo estimates for the probability of selecting alternative models using $\Psi_{CV(m)}(Y)$, with $n = 200$ and $m = 1, 5, 10, 20, 30, 50, 70, 100, 150$. Estimates based on n-dimensional simulated samples from the true linear Gaussian regression model with 5 covariates. The candidate regression models are nested and they have a number $d-2$ of covariates ranging from 1 to 9.

A Appendix

Proof of Proposition 2.1. By means of a stochastic Taylor expansion for $\partial_r \ell(\widehat{\omega}_{(q)}; Y)$, $r = 1, \ldots, d$, around $\widehat{\omega}_{(q)} = \widehat{\omega}$ we have that, up to terms of lower asymptotic order,

$$\partial_r \ell(\widehat{\omega}_{(q)}; Y) \doteq \partial_r \ell(\widehat{\omega}; Y) + (\widehat{\omega}_{(q)s} - \widehat{\omega}_s)\partial_{rs} \ell(\widehat{\omega}; Y).$$

Multiplication of both sides by $\partial^{ru}\ell(\widehat{\omega};Y)$ gives

$$(\widehat{\omega}_{(q)u} - \widehat{\omega}_u) = \partial_r\ell(\widehat{\omega}_{(q)};Y_q)\partial^{ru}\ell(\widehat{\omega};Y) + o_p(n^{\delta-1}), \quad u = 1,\ldots,d, \tag{A.1}$$

with $\partial_r\ell(\widehat{\omega}_{(q)};Y_q) = \sum_{i\in q}\partial_r\ell(\widehat{\omega}_{(q)};Y_i)$. This simplified expression is obtained since $\widehat{\omega}$ and $\widehat{\omega}_{(q)}$ satisfy (2.2) and (2.3), respectively.

Let us consider the following expansion for the term $\partial_r\ell(\widehat{\omega}_{(q)};Y_q)$ in equation (A.1)

$$\partial_r\ell(\widehat{\omega}_{(q)};Y_q) \doteq \partial_r\ell(\widehat{\omega};Y_q) + (\widehat{\omega}_{(q)s} - \widehat{\omega}_s)\partial_{rs}\ell(\widehat{\omega};Y_q), \quad r = 1,\ldots,d,$$

where terms of lower order are not taken into account. Since $\partial_{rs}\ell(\widehat{\omega};Y_q) = \sum_{i\in q}\partial_{rs}\ell(\widehat{\omega};Y_i) = O_p(n^{\delta})$, with $\delta \in (0,1)$, we state that

$$\partial_r\ell(\widehat{\omega}_{(q)};Y_q) = \partial_r\ell(\widehat{\omega};Y_q) + o_p(n^{\delta}) \tag{A.2}$$

and substitution of equation (A.2) in (A.1) completes the proof. □

Proof of Proposition 3.1. Let us start with the proof of (3.2). By expanding $\log f(Z;\widehat{\omega})$ in a Taylor series around $\widehat{\omega} = \omega^*$ we obtain that

$$\log f(Z;\widehat{\omega}) = \log f(Z;\omega^*) + (\widehat{\omega}_r - \omega_r^*)\partial_r\ell(\omega^*;Z) + \frac{1}{2}(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)\partial_{rs}\ell(\omega^*;Z) + O_p(n^{-1/2}).$$

Since $E_Z\{\log f(Z;\omega^*)\}$ is equal to $E_Y\{\log f(Y;\omega^*)\}$ and, by considering (2.4), $E_Z\{\partial_r\ell(\omega^*;Z)\} = 0$, exactly or to the relevant order of approximation, we state that

$$E_Z\{\log f(Z;\widehat{\omega})\} = E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)E_Z\{\partial_{rs}\ell(\omega^*;Z)\} + O_p(n^{-1/2}).$$

Taking expectations term by term, with respect to the true distribution of Y, and using relation (2.5), we get the final expansion

$$\begin{aligned}\eta(g,\widehat{f}) &= E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\nu_{t,u}i^{rt}i^{us}\nu_{rs} + O(n^{-1}) \\ &= E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\nu_{t,r}i^{rt} + O(n^{-1}).\end{aligned}$$

In order to prove (3.3), we derive the following Taylor series around $\widehat{\omega} = \omega^*$

$$\ell(\widehat{\omega};Y) = \ell(\omega^*;Y) + (\widehat{\omega}_r - \omega_r^*)\partial_r\ell(\omega^*;Y) + \frac{1}{2}(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)\partial_{rs}\ell(\omega^*;Y) + O_p(n^{-1/2})$$

and around $\omega^* = \widehat{\omega}$

$$\partial_r\ell(\omega^*;Y) = \partial_r\ell(\widehat{\omega};Y) + (\omega_s^* - \widehat{\omega}_s)\partial_{rs}\ell(\widehat{\omega};Y) + O_p(1).$$

Using the fact that $\partial_{rs}\ell(\widehat{\omega};Y) = \partial_{rs}\ell(\omega^*;Y) + O_p(n^{1/2})$ and that, from (2.2), $\partial_r\ell(\widehat{\omega};Y) = 0$, we obtain

$$\ell(\widehat{\omega};Y) = \ell(\omega^*;Y) - \frac{1}{2}(\widehat{\omega}_r - \omega_r^*)(\widehat{\omega}_s - \omega_s^*)\partial_{rs}\ell(\omega^*;Y) + O_p(n^{-1/2}).$$

Since $(\widehat{\omega}_r - \omega_r^*) = \partial_t\ell(\omega^*;Y)i^{rt} + O_p(n^{-1})$, $r = 1,\ldots,d$, (see Barndorff-Nielsen and Cox, 1994, Section 5.3, under the assumption previously outlined), taking expectations term by term, yields

$$E_Y\{\ell(\widehat{\omega};Y)\} = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\nu_{rs,t,u}i^{rt}i^{su} + O(n^{-1}),$$

with $\nu_{rs,t,u} = E_Y\{\partial_{rs}\ell(\omega^*;Y)\partial_t\ell(\omega^*;Y)\partial_u\ell(\omega^*;Y)\}$. Finally, as a consequence of relation (2.6), we have that $\nu_{rs,t,u} = \nu_{rs}\nu_{t,u} + O(n)$, so that $\nu_{rs,t,u}i^{rt}i^{su} = -\nu_{t,r}i^{rt} + O(n^{-1})$, and this completes the proof. □

Proof of Theorem 3.1. By expanding $\log f(Y_q;\widehat{\omega}_{(q)})$ in a Taylor series around $\widehat{\omega}_{(q)} = \widehat{\omega}$, we obtain the following asymptotic relation

$$\begin{aligned}\Psi_{CV(m)}(Y) &= \ell(\widehat{\omega};Y) + \frac{1}{\binom{n-1}{m-1}}\sum_q(\widehat{\omega}_{(q)r} - \widehat{\omega}_r)\partial_r\ell(\widehat{\omega};Y_q)\\ &+ \frac{1}{2}\frac{1}{\binom{n-1}{m-1}}\sum_q(\widehat{\omega}_{(q)r} - \widehat{\omega}_r)(\widehat{\omega}_{(q)s} - \widehat{\omega}_s)\partial_{rs}\ell(\widehat{\omega};Y_q) + O_p(n^{-1}).\end{aligned}$$

In this case, $\partial_{rs}\ell(\widehat{\omega};Y_q) = O_p(m) = O_p(n^{\delta})$, so that we consider the expansion up to the third term, which is of order $O_p(n^{\delta-1})$. Indeed, since $\delta \neq 1$, relation (2.7), linking $\widehat{\omega}_{(q)}$ and $\widehat{\omega}$, is valid and we state that $\Psi_{CV(m)}(Y)$ corresponds to the following modification of the the maximized loglikelihood

$$\begin{aligned}\Psi_{CV(m)}(Y) &= \ell(\widehat{\omega};Y) + \frac{1}{\binom{n-1}{m-1}}\sum_q\{\partial_r\ell(\widehat{\omega};Y_q)\partial_s\ell(\widehat{\omega};Y_q)\}\,\partial^{rs}\ell(\widehat{\omega};Y)\\ &+ \frac{1}{2}\frac{1}{\binom{n-1}{m-1}}\sum_q\{\partial_t\ell(\widehat{\omega};Y_q)\partial_u\ell(\widehat{\omega};Y_q)\partial_{rs}\ell(\widehat{\omega};Y_q)\}\,\partial_{tr}\ell(\widehat{\omega};Y)\partial_{us}\ell(\widehat{\omega};Y) + O_p(n^{-1}). \quad \text{(A.3)}\end{aligned}$$

The associated expected value is, using expansion (3.3),

$$\begin{aligned}E_Y\{\Psi_{CV(m)}(Y)\} &= E_Y\{\log f(Y;\omega^*)\} + \frac{1}{2}\nu_{t,r}i^{rt}\\ &+ \frac{1}{\binom{n-1}{m-1}}\sum_q E_Y\{\partial_r\ell(\omega^*;Y_q)\partial_s\ell(\omega^*;Y_q)\partial^{rs}\ell(\omega^*;Y)\}\\ &+ \frac{1}{2}\frac{1}{\binom{n-1}{m-1}}\sum_q E_Y\{\partial_t\ell(\omega^*;Y_q)\partial_u\ell(\omega^*;Y_q)\partial_{rs}\ell(\omega^*;Y_q)\partial_{tr}\ell(\omega^*;Y)\partial_{us}\ell(\omega^*;Y)\}\\ &+ O(n^{-1}),\end{aligned}$$

where ω^* is substituted for $\widehat{\omega}$ in the second and in the third terms of the right hand side of (A.3). As a consequence of (2.6), we have that $\partial^{rs}\ell(\omega^*;Y) = -i^{rs} + O_p(n^{-3/2})$. Finally, since $\delta \neq 0$, a

relation analogous to (2.6) holds for $\partial_{rs}\ell(\omega^*;Y_q)$, so that $E_Y\{\partial_t\ell(\omega^*;Y_q)\partial_u\ell(\omega^*;Y_q)\partial_{rs}\ell(\omega^*;Y_q)\} = \nu_{q;rs}\nu_{q;t,u} + O(n^\delta)$ and this completes the proof. □

Proof of Theorem 3.2. By expanding $\log f(Y_q;\widehat{\omega}_{(q)})$ in a Taylor series around $\widehat{\omega}_{(q)} = \omega^*$, we have that

$$\begin{aligned}\Psi_{CV(m)}(Y) &= \log f(Y;\omega^*) + \frac{1}{\binom{n-1}{m-1}}\sum_q(\widehat{\omega}_{(q)r} - \omega^*_r)\partial_r\ell(\omega^*;Y_q) \\ &+ \frac{1}{2}\frac{1}{\binom{n-1}{m-1}}\sum_q(\widehat{\omega}_{(q)r} - \omega^*_r)(\widehat{\omega}_{(q)s} - \omega^*_s)\partial_{rs}\ell(\omega^*;Y_q) + O_p(n^{1-3\gamma/2}). \qquad (A.4)\end{aligned}$$

Since m increases with n, that is $\delta \neq 0$, we have that $E_{Y_q}\{\partial_r\ell(\omega^*;Y_q)\} = 0$, exactly or to the relevant order of approximation. Indeed, a result analogous to (2.5) may be considered for $\widehat{\omega}_{(q)}$ and then, taking the expectation terms by terms in (A.4), with respect to the true density of Y, we obtain relation (3.8). □

Proof of Corollary 5.1. For a Gaussian regression model we have that $\partial_{ru}\ell(\widehat{\omega};Y) = \partial_{r\sigma^2}\ell(\widehat{\omega};Y) = 0$, $r = 1,\ldots,d-1$, $u = d$, $\partial_{ru}\ell(\widehat{\omega};Y) = \partial_{\sigma^2\sigma^2}\ell(\widehat{\omega};Y) = -n/(2\widehat{\sigma}^4)$, $r,u = d$, and

$$\partial_{ru}\ell(\widehat{\omega};Y) = -\frac{1}{\widehat{\sigma}^2}\left[\sum_{i=1}^n \frac{x_{ir}x_{iu}}{\{\eta'(\widehat{\mu}_i)\}^2} + \sum_{i=1}^n \frac{(Y_i - \widehat{\mu}_i)x_{ir}x_{iu}\eta''(\widehat{\mu}_i)}{\{\eta'(\widehat{\mu}_i)\}^3}\right],\ r,u = 1,\ldots,d-1, \qquad (A.5)$$

with $\eta''(\cdot)$ the second derivative of $\eta(\cdot)$. Substitution in (2.7) completes the proof. □

Proof of Proposition 5.1. Equation (5.1) derives from (3.2), since for a linear Gaussian regression model it is easy to see that $E_Y\{\log f(Y;\omega_0)\} = -n\left\{\log(2\pi\sigma_0^2) + 1\right\}/2$. Furthermore, $d \geq d_0$ assures that the model contains the true distribution, so that $\nu_{t,r}i^{rt} = d$.

Moreover, since in this case

$$\nu_{q;r,s} = \frac{\sum_{i\in q}x_{ir}x_{is}}{\sigma_0^2},\ r,s = 1,\ldots,d-1, \quad \nu_{q;r,\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad \nu_{q;\sigma^2,\sigma^2} = \frac{m}{2\sigma_0^4},$$

$$\nu_{q;rs} = -\frac{\sum_{i\in q}x_{ir}x_{is}}{\sigma_0^2},\ r,s = 1,\ldots,d-1, \quad \nu_{q;r\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad \nu_{q;\sigma^2\sigma^2} = -\frac{m}{2\sigma_0^4},$$

$$i^{rs} = \sigma_0^2\Sigma_{rs} + O(n^{-2}),\ r,s = 1,\ldots,d-1, \quad i^{r\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad i^{\sigma^2\sigma^2} = \frac{2\sigma_0^4}{n},$$

using equation (3.5), we prove that the expected value of $\Psi_{CV(m)}(Y)$, with $m = O(n^\delta)$ and $n - m = O(n^\gamma)$, $\delta \in (0,1)$, $\gamma \in (0,1]$, corresponds to (5.2).

Finally, in order to compute the mean value of $\Psi_{CV(m)}(Y)$ for the case when $m = O(n)$ and $n - m = O(n^\gamma)$, $\gamma \in (1/2, 1]$, we have to use the alternative expansion (3.8) and to consider that

$$\nu_{(q);r,s} = \frac{\sum_{i \notin q} x_{ir} x_{is}}{\sigma_0^2},\ r,s = 1,\ldots,d-1, \quad \nu_{(q);r,\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad \nu_{(q);\sigma^2,\sigma^2} = \frac{n-m}{2\sigma_0^4},$$

$$i_{(q)}^{rs} = \sigma_0^2 \Sigma_{rs} + O(n^{-2\gamma}),\ r,s = 1,\ldots,d-1, \quad i_{(q)}^{r\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad i_{(q)}^{\sigma^2\sigma^2} = \frac{2\sigma_0^4}{n-m}.$$

□

Proof of Proposition 5.2. Since

$$i^{rs} = \frac{\sigma^{*2}\Sigma_{rs}}{n} + O(n^{-2}),\ r,s = 1,\ldots,d-1, \quad i^{r\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad i^{\sigma^2\sigma^2} = \frac{2\sigma^{*4}}{n},$$

$$\nu_{r,s} = \frac{\sigma_0^2 \sum_{i=1}^n x_{ir} x_{is}}{\sigma^{*4}},\ r,s = 1,\ldots,d-1, \quad \nu_{\sigma^2,\sigma^2} = \frac{n\sigma_0^4}{2\sigma^{*8}} + \frac{\sigma_0^4 \rho}{\sigma^{*8}} + O(1),$$

using (3.2), we find that the target mean value is

$$\eta(g, \widehat{f}) = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\left(\frac{\sigma_0^2 \sum_{i=1}^n x_{ir} x_{is} \Sigma_{rs}}{n\sigma^{*2}} + \frac{\sigma_0^4}{\sigma^{*4}} + \frac{2\sigma_0^4 \rho}{\sigma^{*4} n}\right) + O(n^{-1}),$$

which, after simple calculations, gives (5.4).

With regard to the expectation of $\Psi_{CV(m)}(Y)$, for $\delta = 0$, $\gamma = 1$, we prove that, being m fixed,

$$\nu_{q;r,s} = \frac{\sigma_0^2 \sum_{i \in q} x_{ir} x_{is}}{\sigma^{*4}} + \frac{1}{\sigma^{*4}} \sum_{i \in q} \sum_{j \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)(\beta_0^T x_{0j} - \beta^{*T} x_j) x_{ir} x_{ijs},\ r,s = 1,\ldots,d-1,$$

$$\begin{aligned} \nu_{q;\sigma^2,\sigma^2} &= \frac{m\sigma_0^4}{2\sigma^{*8}} + \frac{\sigma_0^2}{\sigma^{*8}} \sum_{i \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)^2 + \frac{1}{4\sigma^{*8}} \sum_{i \in q} \sum_{j \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)^2 (\beta_0^T x_{0j} - \beta^{*T} x_j)^2 \\ &+ \frac{m}{2\sigma^{*6}} \left(\frac{\sigma_0^2}{\sigma^{*2}} - 1\right) \sum_{i \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)^2 + \frac{m^2}{4\sigma^{*4}} \left(\frac{\sigma_0^4}{\sigma^{*4}} - 2\frac{\sigma_0^2}{\sigma^{*2}} + 1\right). \end{aligned}$$

Moreover,

$$\frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{q;r,s} = \nu_{r,s} + \frac{1}{\sigma^{*4}\binom{n-1}{m-1}} \sum_q \sum_{i \in q} \sum_{j \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)(\beta_0^T x_{0j} - \beta^{*T} x_j) x_{ir} x_{js},\ r,s = 1,\ldots,d-1,$$

$$\begin{aligned} \frac{1}{\binom{n-1}{m-1}} \sum_q \nu_{q;\sigma^2,\sigma^2} &= \nu_{\sigma^2,\sigma^2} + \frac{1}{4\sigma^{*8}\binom{n-1}{m-1}} \sum_q \sum_{i \in q} \sum_{j \in q} (\beta_0^T x_{0i} - \beta^{*T} x_i)^2 (\beta_0^T x_{0j} - \beta^{*T} x_j)^2 \\ &+ \frac{m}{2\sigma^{*6}} \left(\frac{\sigma_0^2}{\sigma^{*2}} - 1\right) \sum_{i=1}^n (\beta_0^T x_{0i} - \beta^{*T} x_i)^2 + \frac{nm}{4\sigma^{*4}} \left(\frac{\sigma_0^4}{\sigma^{*4}} - 2\frac{\sigma_0^2}{\sigma^{*2}} + 1\right) \end{aligned}$$

and, using formula (3.6), we obtain

$$E_Y\{\Psi_{CV(m)}(Y)\} = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\left\{\frac{d+2(d-1)\rho/n}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O(n^{-1}),$$

which corresponds to (5.5).

Finally, when $\delta \in (0,1]$ and $\gamma \in (1/2,1]$, we have that $\sigma^{*2} = \sigma_0^2\{1+\rho/(n-m)\} + O(n^{-\gamma})$, where $\rho = \sigma_0^{-2}\sum_{i\notin q}(\beta_0^T x_{0i} - \beta_{n-m}^{*T} x_i)^2$, and $\rho/(n-m)$ equals, neglecting the additional terms of order $O(n^{-\gamma})$, the quantity ρ/n . Indeed, since both m and $n-m$ increase with n,

$$i_{(q)}^{rs} = \frac{\sigma^{*2}\Sigma_{rs}}{n-m} + O(n^{-2\gamma}),\ r,s = 1,\ldots,d-1, \quad i_{(q)}^{r\sigma^2} = 0,\ r = 1,\ldots,d-1, \quad i_{(q)}^{\sigma^2\sigma^2} = \frac{2\sigma^{*4}}{n-m},$$

$$\nu_{(q);r,s} = \frac{\sigma_0^2\sum_{i\notin q} x_{ir}x_{is}}{\sigma^{*4}},\ r,s = 1,\ldots,d-1, \quad \nu_{(q);\sigma^2,\sigma^2} = \frac{(n-m)\sigma_0^4}{2\sigma^{*8}} + \frac{\sigma_0^4\rho}{\sigma^{*8}} + O(1),$$

$$\nu_{q;rs} = -\frac{\sum_{i\in q} x_{ir}x_{is}}{\sigma^{*2}},\ r,s = 1,\ldots,d-1, \quad \nu_{q;\sigma^2\sigma^2} = -\frac{m}{2\sigma^{*4}} + O(1).$$

Thus, it is immediate to see that

$$\begin{aligned}
\frac{1}{2}\frac{1}{\binom{n-1}{m-1}}\sum_q \nu_{(q);t,u}\, i_{(q)}^{rt}\, i_{(q)}^{us}\, \nu_{q;rs} &= -\frac{1}{2}\frac{m}{(n-m)}\frac{1}{\binom{n-1}{m-1}}\sum_q \nu_{(q);t,r}\, i_{(q)}^{rt} \\
&= -\frac{1}{2}\frac{m}{(n-m)}\frac{\binom{n}{m}}{\binom{n-1}{m-1}}\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O(n^{1-2\gamma})
\end{aligned}$$

and, using (3.8), we find that

$$E_Y\{\Psi_{CV(m)}(Y)\} = E_Y\{\log f(Y;\omega^*)\} - \frac{1}{2}\frac{n}{(n-m)}\left\{\frac{d}{1+\rho/n} + \frac{\rho/n}{(1+\rho/n)^2}\right\} + O(n^{1-2\gamma}),$$

which equals (5.6).

□

References

[1] Allen, D.M. (1974), The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, 16, 125–127.

[2] Akaike, H. (1973), *Information theory and extension of the maximum likelihood principle*, in: N.B. Petron and F. Caski (Eds.), Second Symposium on Information Theory, Akademiai Kiado, Budapest, 267–281 .

[3] Arlot, S. (2008), V-fold cross-validation improved: V-fold penalization, arXiv: 0802.0566v2.

[4] Arlot, S. and A. Celisse (2010), A survey of cross-validation procedures for model selection, Statistics Surveys, 4, 40–79.

[5] Barndorff-Nielsen, O.E. and D.R. Cox (1994), *Inference and Asymptotics*, Chapman and Hall, London.

[6] Burman, P. (1989), A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods, Biometrika, 76, 503–514.

[7] Burnham, K.P. and D.R. Anderson (2002), *Model Selection and Multimodel Inference*. (2nd edition), Springer-Verlag, New York.

[8] Celisse, A. (2014), Optimal cross-validation in density estimation with the L^2-loss, Annals of Statistics, 42, 1879–1910.

[9] Davison, A.C. and D.V. Hinkley (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

[10] Draper, N.R. and H. Smith (1998), *Applied Regression Analysis*, (3rd edition), John Wiley, New York.

[11] Efron, B. (1983), Estimating the error rate of a prediction rule: improvement of cross-validation, Journal of the American Statistical Association, 78, 316–331.

[12] Efron, B. (1986), How biased is the apparent error rate of a predition rule?, Journal of the American Statistical Association, 81, 461–470.

[13] Fujikoshi, Y., T. Noguchi, M. Ohtaki and H. Yanagihara (2003), Corrected versions of cross-validation criteria for selecting multivariate regression and growth models, Annals of the Institute of Statistical Mathematics, 55, 537–553.

[14] Fushiki, T. (2011), Estimation of prediction error by using K-fold cross-validation, Statistics and Computing, 21, 137–146.

[15] Geisser, S. (1975), The predictive sample reuse method with applications, Journal of the American Statistical Association, 70, 320–328.

[16] Geisser, S. and W.F. Eddy (1979), A predictive approach to model selection, Journal of the American Statistical Association, 74, 153–160.

[17] Herzberg, G. and S. Tsukanov (1986), A note on modifications of the jacknife criterion on model selection, Utilitas Mathematica, 29, 209–216.

[18] Konishi, S. and G. Kitagawa (1996), Generalised information criteria in model selection, Biometrika, 83, 875–890.

[19] Li, K.-C. (1987), Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set, Annals of Statistics, 15, 958–975.

[20] Picard, R.R. and R.D. Cook (1984), Cross-validation of regression models, Journal of the American Statistical Association, 79, 575–583.

[21] Shao, J. (1993), Linear model selection by cross-validation, Journal of the American Statistical Association, 88, 486–494.

[22] Shao, J. (1997), An asymptotic theory for linear model selection, Statistica Sinica, 7, 221-264.

[23] Stone, M. (1974), Cross-validation choice and assessment of statistical predictions, Journal of the Royal Statistical Society: Series B, 36, 111–147.

[24] Stone, M. (1977a), An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, Journal of the Royal Statistical Society: Series B, 39, 44–47.

[25] Stone, M. (1977b), Asymptotics for and against cross-validation, Biometrika, 64, 29–38.

[26] White, H. (1994), *Estimation, Inference, and Specification Analysis*, Cambridge University Press, New York.

[27] Zhang, P. (1992), On the distributional properties of model selection criteria, Journal of the American Statistical Association, 87, 732–737.

[28] Yanagihara, H. and H. Fujisawa (2012), Iterative bias correction of the cross-validation criterion, Scandinavian Journal of Statistics, 39, 116–130.

[29] Yanagihara, H., T. Tonda and C. Matsumoto (2006), Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition, Journal of Multivariate Analysis, 97, 1965–1975.

[30] Yang, Y. (2005), Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, Biometrika, 92, 937–950.

[31] Yang, Y. (2007), Consistency of cross validation for comparing regression procedures, Annals of Statistics, 35, 2450–2473.

[32] Zhang, P. (1993), Model selection via multifold cross-validation, Annals of Statistics, 21, 299–313.